Estimation and Forecasting in Vector Autoregressive Moving Average Models for Rich Datasets

Gustavo Fruet Dias*
Department of Economics and Business Economics, Aarhus University
CREATES

George Kapetanios
King's College London

This version: May, 2017

Abstract

We address the issue of modelling and forecasting macroeconomic variables using rich datasets by adopting the class of Vector Autoregressive Moving Average (VARMA) models. We overcome the estimation issue that arises with this class of models by implementing an iterative ordinary least squares (IOLS) estimator. We establish the consistency and asymptotic distribution of the estimator for weak and strong VARMA(p,q) models. Monte Carlo results show that IOLS is consistent and feasible for large systems, outperforming the MLE and other linear regression based efficient estimators under alternative scenarios. Our empirical application shows that VARMA models are feasible alternatives when forecasting with many predictors. We show that VARMA models outperform the AR(1), ARMA(1,1), Bayesian VAR, and factor models, considering different model dimensions.

JEL classification numbers: C13, C32, C53, C63, E0

Keywords: VARMA, weak VARMA, Iterative ordinary least squares (IOLS) estimator, Asymptotic contraction mapping, Forecasting, Rich and Large datasets.

*Corresponding author at: Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark. E-mail: gdias@econ.au.dk.

1 Introduction

The use of large arrays of economic indicators to forecast key macroeconomic variables has become very popular recently. Economic agents consider a wide range of information when they construct their expectations about the behaviour of macroeconomic variables such as interest rates, industrial production, and inflation. In the past several years, this information has become more widely available through a large number of indicators that aim to describe different sectors and fundamentals from the whole economy. To improve forecast accuracy, large sized datasets that attempt to replicate the set of information used by agents to make their decisions are incorporated into econometric models.

For the past twenty years, macroeconomic variables have been forecasted using vector autoregression (VAR) models. This type of model performs well when the number of variables in the system is relatively small. When the number of variables increases, however, the performance of VAR forecasts deteriorates very fast, generating the so-called "curse of dimensionality". In this paper, we propose the use of vector autoregressive moving average (VARMA) models, estimated with the iterative ordinary least squares (IOLS) estimator, as a feasible method to address the "curse of dimensionality" on medium and large sized datasets and improve forecast accuracy of macroeconomic variables. To the best of our knowledge, the VARMA methodology has never been applied to medium and large sized datasets, as we do in this paper.

VARMA models have been studied for the past thirty years, but they have never been as popular as VAR models because of estimation and specification issues. Despite having attractive theoretical properties, estimation of VARMA models remains a challenge. Linear estimators (Hannan and Rissanen (1982), Hannan and Kavalieris (1984), Dufour and Jouini (2014), among others) and Bayesian methods (Chan et al. (2016)) have been proposed in the literature as a way to overcome the numerical difficulties posed by the efficient maximum-likelihood estimator. We address the estimation issue and hence contribute to this literature by proposing the use of the IOLS estimator that is feasible for high-dimensional VARMA models.

Other methodologies have been proposed in the literature to deal with the "curse of dimensionality". We can divide them mainly in two groups of models. The first aims to overcome the dimensionality issue by imposing restrictions on the parameter matrices of a standard VAR model. Among the many important contributions from this field, we

point out the following classes of models: Bayesian VAR (BVAR) (De Mol et al. (2008) and Bańbura et al. (2010)); Ridge (De Mol et al. (2008)); reduced rank VAR (Carriero et al. (2011)); and Lasso (De Mol et al. (2008) and Tibshirani (1996)). The second group of models dealing with the "curse of dimensionality" reduces the dimension of the dataset by constructing summary proxies from the large dataset. Chief among these models is the class of factor models. The seminal works in this area are Forni et al. (2000) and Stock and Watson (2002a,b). Common factor models improve forecast accuracy and produce theoretically well-behaved impulse response functions, as reported by De Mol et al. (2008) and Bernanke et al. (2005).

VARMA models are able to capture two important features from these two groups of models. The first is the reduction of the model dimensionality, achieved by setting some elements of the parameter matrices to zero following uniqueness requirements. This produces parameter matrices which are not full rank, resembling the reduced rank VAR model of Carriero et al. (2011). Furthermore, VARMA models parsimoniously account for sample correlation profiles of different shapes than the geometrically declining sinusoids associated with the VAR models. The second is the relationship between VARMA models and factor representations of a vector stochastic process. If the latent common factors follow either a finite VAR or VARMA processes, then the observed series will have a VARMA representation (Dufour and Stevanović (2013)). This fact reinforces the use of VARMA models as a suitable framework to forecast key macroeconomic variables using potentially many predictors. Additionally, VARMA models are closed under linear transformation and marginalization (see Lütkepohl (2007, Section 11.6.1)), providing additional flexibility and potentially better forecast performance.

With regard to the theory, we establish the consistency and asymptotic distribution of the IOLS estimator for the weak and strong VARMA(p,q) models. Strong VARMA models have innovations that are independently identically distributed (*i.i.d.*), while the disturbances of the weak VARMA models are only uncorrelated. Our asymptotic results are obtained under mild assumptions using the asymptotic contraction mapping framework of Dominitz and Sherman (2005). We conduct an extensive Monte Carlo study, considering different system dimensions, sample sizes, weak and strong innovations, eigenvalues associated with the parameter matrices, and Kronecker indices specifications. We show that the IOLS estimator has a good finite sample performance when compared to alternative esti-

mators, such as the efficient maximum-likelihood (MLE) estimator, and important linear competitors. Specifically, the IOLS estimator performs well in a variety of scenarios such as small sample size; the eigenvalues associated with the parameter matrices are near-to-zero; and high-dimensional systems.

In the empirical part of the paper, we focus on forecasting three key macroeconomic variables: industrial production, interest rate, and CPI inflation using potentially large sized datasets. As in Carriero et al. (2011), we use the 52 US macroeconomic variables taken from the dataset provided by Stock and Watson (2006) to construct systems with five different dimensions: 3, 10, 20, 40, and 52 variables. By doing that, we are able to evaluate the tradeoff between forecast gains by incorporating more information (large sized datasets) and the estimation cost associated with it. Additionally, the different system dimensions play the role of robustness check. We show that VARMA models are strong competitors and produce more accurate forecasts than the benchmark models (AR(1), ARMA(1,1), BVAR, and factor models) in different occasions. This conclusion holds for different system sizes and horizons.

The paper is structured as follows. In Section 2, we discuss the properties and identification of VARMA models and derive the IOLS estimator. In Section 3, we establish the consistency and asymptotic distribution of the IOLS estimator considering the general weak and strong VARMA(p,q) models. In Section 4, we address the consistency and efficiency of VARMA models estimated with the IOLS procedure through a Monte Carlo study. In Section 5, we display the results of our empirical application. The proofs and tables are relegated to an Appendix. An online Supplement presents proofs for the auxiliary Lemmas, the entire set of tables with results for the Monte Carlo and the empirical studies, and further discussion on selected topics.

2 VARMA Models, Identification, and Estimation Procedures

Our interest lies in modelling and forecasting key elements of the K-dimensional vector process $Y_t = (y_{1,t}, y_{2,t}, ..., y_{K,t})'$, where K is allowed to be large. We assume, as a baseline

model, a general nonstandard VARMA(p,q) model where the means have been removed,[1]

$$A_0 Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + ... + A_p Y_{t-p} + M_0 U_t + M_1 U_{t-1} + ... + M_q U_{t-q}, \quad (1)$$

where the disturbances $U_t = (u_{1,t}, u_{2,t}, ..., u_{K,t})'$ are assumed to be a zero-mean sequence of uncorrelated random variables (weak white noise process) with a nonsingular $(K \times K)$ covariance matrix Σ_u, and Y has dimension $(K \times 1)$. Our baseline model is assumed to be stable and invertible.[2]

2.1 Identification and Uniqueness

The issue of identifying unique parameterizations of VARMA models has been an important topic of study in econometrics and statistics (Hannan (1969, 1976), Akaike (1974, 1976), Hannan and Kavalieris (1984), Hannan and Deistler (1988), among others). This follows because nonstandard VARMA models require restrictions on the parameter matrices to ensure that the model is uniquely identified. To formally define uniqueness of a VARMA representation, define the lag polynomials $A(L) = A_0 - A_1 L - A_2 L^2 - ... - A_p L^p$ and $M(L) = M_0 + M_1 L + M_2 L^2 + ... + M_q L^q$, where L is the usual lag operator. More specifically, we say that a model is unique if there is only one pair of stable and invertible polynomials $A(L)$ and $M(L)$, respectively, which satisfies the canonical MA representation

$$Y_t = A(L)^{-1} M(L) = \Theta(L) U_t = \sum_{i=0}^{\infty} \Theta_i U_{t-i}, \quad (2)$$

for a given $\Theta(L)$ operator. In contrast to the reduced form VAR models, setting $A_0 = M_0 = I_K$ is not sufficient to ensure a unique VARMA representation. Uniqueness of (1) is guaranteed by imposing restrictions on the $A(L)$ and $M(L)$ operators in that these operators are unique left-coprime, i.e. the only feasible left divisor of $[A(L) : M(L)]$ is the unimodular operator given by the identity matrix.[3]

A number of different strategies can be implemented to obtain unique VARMA representations, such as the extended scalar component approach of Athanasopoulos and Vahid

[1] We adopt the terminology in Lütkepohl, 2007, p. 448 and call a VARMA representation nonstandard when A_0 and M_0 are allowed to be nonidentity invertible matrices. If $A_0 = M_0 = I_K$, we call it a standard VARMA model.

[2] A general VARMA(p,q) is considered stable and invertible if $\det(A_0 - A_1 z - A_2 z^2 - ... - A_p z^p) \neq 0$ for $|z| \leq 1$ and $\det(M_0 + M_1 z + M_2 z^2 + ... + M_q z^q) \neq 0$ for $|z| \leq 1$ hold, respectively.

[3] $C(L)$ is an unimodular operator if $\det(C(L))$ is a nonzero constant that does not depend on L.

(2008), the final equations form (see Lütkepohl, 2007, pg.362) and the Echelon form transformation (Hannan and Kavalieris (1984), Poskitt (1992), Lütkepohl and Poskitt (1996), among others). Although Athanasopoulos et al. (2012) show that scalar components perform slightly better than the Echelon form methodology in empirical exercises, the authors argue that the latter has the advantage of having a simpler identification procedure. More specifically, the Echelon form identification strategy can be fully automated and provides a more parsimonious parametrization, when compared to final equations. These are highly desired features when modelling medium and large systems. In this paper, we implement the Echelon form transformation as a way to impose uniqueness in both Monte Carlo and empirical applications. Nevertheless, because the three identification strategies (scalar components, final equations form, and Echelon form) impose uniqueness through a set of linear restrictions on the $A(L)$ and $M(L)$ operators, the IOLS estimator can be directly implemented no matter which identification strategy the researcher chooses.

A general VARMA model such as the one stated in (1) is considered to be in its Echelon form if the conditions stated in equations (3), (4), (5), and (6) are satisfied (see Lütkepohl, 2007, p. 452 and Lütkepohl and Poskitt (1996) for more details):

$$p_{ki} := \begin{cases} \min\left(p_k + 1, p_i\right) \text{ for } k \geq i \\ \min\left(p_k, p_i\right) \text{ for } k < i, \end{cases} \qquad \text{for } k, i = 1, ..., K, \tag{3}$$

$$\alpha_{kk}(L) = 1 - \sum_{j=1}^{p_k} \alpha_{kk,j} L^j \qquad \text{for } k = 1, ..., K, \tag{4}$$

$$\alpha_{ki}(L) = - \sum_{j=p_k - p_{ki}+1}^{p_k} \alpha_{ki,j} L^j \qquad \text{for } k \neq i, \tag{5}$$

$$m_{ki}(L) = \sum_{j=0}^{p_k} m_{ki,j} L^j, \qquad \text{for } k, i = 1, ..., K \qquad \text{with} \quad M_0 = A_0, \tag{6}$$

where $A(L)=[\alpha_{ki}]_{k,i=1,...,K}$ and $M(L)=[m_{ki}]_{k,i=1,...,K}$ are, respectively, the operators from the autoregressive and moving average components of the VARMA process; the p_{ki} numbers specify the free coefficients in the operator $\alpha_{ki}(L)$ for $i \neq k$ from the $A(L)$ polynomial; and the arguments p_k with $k = 1, ..., K$ are Kronecker indices which specify the maximum degrees in each row of both $A(L)$ and $M(L)$ polynomials. We define $\mathbf{p} = (p_1, p_2, ..., p_K)'$ as the vector collecting the Kronecker indices (Section 5.1 discusses alternative procedures to choose the Kronecker indices).

The Echelon form formulae in (3) to (6) deliver the necessary and sufficient restrictions for the unique identification of the VARMA models (Hannan and Deistler, 1988, Chapter 2). Notably, the Echelon representation allows the baseline model in (1) to be rewritten as

$$Y_t = (I_K - A_0)(Y_t - U_t) + A_1 Y_{t-1} + A_2 Y_{t-2} + ... + A_p Y_{t-p} + U_t + M_1 U_{t-1} + ... + M_q U_{t-q}, \quad (7)$$

where $(Y_t - U_t) = A_0^{-1} \{\sum_{i=1}^{p} A_i Y_{t-i} + \sum_{i=1}^{q} M_i U_{t-i}\}$ is uncorrelated with U_t. A compact notation of (7) reads

$$\text{vec}(Y) = \left(X' \otimes I_K\right) \text{vec}(B) + \text{vec}(U), \quad (8)$$

where $B = [(I_K - A_0), A_1, ..., A_p, M_1, ..., M_q]$ with dimension $(K \times K(p+q+1))$; $X = (X_{\bar{q}+1}, ..., X_T)$ is the matrix of regressors with dimension $(K(p+q+1) \times T - \bar{q})$, where $\bar{q} = \max\{p, q\}$ and $X_t = \text{vec}(Y_t - U_t, Y_{t-1}, ..., Y_{t-p}, U_{t-1}, ..., U_{t-q})$; $Y = (Y_{\bar{q}+1}, ..., Y_T)$ has dimension $(K \times T - \bar{q})$; and $U = (U_{\bar{q}+1}, ..., U_T)$ is a $(K \times T - \bar{q})$ matrix of disturbances. Note that by setting the dimensions of the X and Y matrices as $(K(p+q+1) \times T - \bar{q})$ and $(K \times T - \bar{q})$, respectively, we explicitly highlight the fact that we lose $\bar{q}$ observations in finite sample. We obtain the free parameters in the model (excluding the distinct covariance parameters) by rewriting $\text{vec}(B)$ into the product of a $\left(K^2(p+q+1) \times n\right)$ deterministic matrix R and a $(n \times 1)$ vector β,

$$\text{vec}(Y) = \left(X' \otimes I_K\right) R\beta + \text{vec}(U), \quad (9)$$

where n denotes the number of free parameters in B, and the $(n \times 1)$ vector β concatenates these parameters. The Echelon restrictions imply a unique full-rank column matrix R and, provided that Σ_u is nonsingular, ensure that $rank\{R' [\mathbb{E}(X_t X_t') \otimes I_K] R\} = n$ (Dufour and Jouini (2005, 2014)). It is still possible to impose additional zero restrictions (data dependent) while keeping the model uniquely identified and hence obtain a more parsimonious model (see Lütkepohl and Poskitt, 1996, p. 73). This follows because a Kronecker index only gives the maximum row degree, and hence not all operators in the k^{th} row of $[A(L) : M(L)]$ need to have the same degree as p_k. As an example, consider a VARMA(1,1) model with $K = 3$ and $\mathbf{p} = (1, 0, 0)'$. The resulting VARMA representation expressed in

Echelon form reads

$$\begin{pmatrix} 1 & 0 & 0 \\ a_{21,0} & 1 & 0 \\ a_{31,0} & 0 & 1 \end{pmatrix} Y_t = \begin{pmatrix} a_{11,1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} Y_{t-1} + \begin{pmatrix} 1 & 0 & 0 \\ a_{21,0} & 1 & 0 \\ a_{31,0} & 0 & 1 \end{pmatrix} U_t + \begin{pmatrix} m_{11,1} & m_{12,1} & m_{13,1} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} U_{t-1}. \quad (10)$$

While the zero restrictions imposed in (10) are those required to meet the restriction imposed by the canonical structure, further restrictions that simplify the model could be added (e.g. $a_{21,0} = a_{31,0} = m_{12,1} = m_{13,1} = 0$) without affecting the uniqueness of the VARMA representation. Throughout this entire study, we refrain from adding these additional restrictions and restrict ourselves to the ones implied from (3) to (6). Furthermore, we always consider the Echelon forms that yield VARMA models that cannot be partitioned into smaller independent systems.

Finally, it is important to note that the VARMA representation in (1) is not a structural VARMA (SVARMA) model in its classical definition (see Gourieroux and Monfort (2015)), because (1) is not necessarily driven by independent (or uncorrelated) shocks. To construct impulse response functions, which depend on the structural shocks, additional identification restrictions to the ones required for uniqueness are necessary. In particular, as noted by Gourieroux and Monfort (2015), the structural shocks can usually be derived by imposing restrictions either on the contemporaneous correlation among the innovations in (1) (see the so-called B-model in Lütkepohl, 2007, p. 362 and the general identification theory in Rubio-Ramírez et al. (2010)), or on the long-run impact matrix of the shocks (Blanchard and Quah (1989)), or by imposing sign restrictions on some impulse response functions (Uhlig (2005)). This identification issue is common to both VARMA and VAR models and has been primarily explored in the context of VAR models.

2.2 Estimation

VARMA models, similar to their univariate (ARMA model) counterparts, are usually estimated using MLE procedures. Provided that the model in (1) is uniquely identified and disturbances U_t are normally distributed, MLE delivers consistent and efficient estimators. Although MLE seems to be very powerful at first glance, it presents serious problems when dealing with medium and large sized datasets. We overcome this issue by implementing an

IOLS procedure in the spirit of Spliid (1983), Hannan and McDougall (1988) and Kapetanios (2003).

The IOLS framework consists of computing ordinary least squares (OLS) estimates of the parameters using estimates of the latent regressors. These regressors are computed recursively at each iteration using the OLS estimates. Using the invertibility condition, we can express a finite VARMA model as an infinite VAR, $Y_t = \sum_{i=1}^{\infty} \Pi_i Y_{t-i} + U_t$. We compute consistent estimates of U_t, denoted as $\widehat{U}_t^0$, by truncating the infinite VAR representation using some lag order, $\tilde{p}$, that minimizes some criterion. Following the results in Ng and Perron (1995) and Dufour and Jouini (2005), choosing $\tilde{p}$ proportional to $\ln(T)$ delivers consistent estimates of U_t, so that $\widehat{U}_t^0 = Y_t - \sum_{i=1}^{\tilde{p}} \widehat{\Pi}_i Y_{t-i}$.[4] Substitute $\widehat{U}_t^0$ into the matrix X in (9) and denote it $\widehat{X}^0$. Note that both $\widehat{X}^0$ and $\text{vec}(Y)$ change their dimensions to $(K(p+q+1) \times T - \bar{q} - \tilde{p})$ and $(K(T - \bar{q} - \tilde{p}) \times 1)$, respectively. This happens exclusively on this first iteration, because $\tilde{p}$ observations are lost on the VAR($\tilde{p}$) approximation of U_t. The first iteration of the IOLS estimator is obtained by computing the OLS estimator from the modified version of (9),

$$\widehat{\beta}^1 = \left[R' \left(\widehat{X}^0 \widehat{X}^{0\prime} \otimes I_K\right) R\right]^{-1} R' \left(\widehat{X}^0 \otimes I_K\right) \text{vec}(Y). \tag{11}$$

It is relevant to highlight that the first step of the IOLS estimator is the two-stage Hannan-Rissanen (HR) algorithm formulated in Hannan and Rissanen (1982), and for which Dufour and Jouini (2005) show the consistency and asymptotic distribution.

We are now in a position to use $\widehat{\beta}^1$ to recover the parameter matrices $\widehat{A}_0^1, ..., \widehat{A}_p^1$, $\widehat{M}_1^1, ..., \widehat{M}_q^1$ and a new set of residuals $\widehat{U}^1 = \left(\widehat{U}_1^1, \widehat{U}_2^1, ..., \widehat{U}_T^1\right)$ by recursively applying

$$\widehat{U}_t^1 = Y_t - \left[\widehat{A}_0^1\right]^{-1} \left[\widehat{A}_1^1 Y_{t-1} - ... - \widehat{A}_p^1 Y_{t-p} - \widehat{M}_1^1 \widehat{U}_{t-1}^1 - ... - \widehat{M}_q^1 \widehat{U}_{t-q}^1\right], \text{ for } t = 1, ..., T, \tag{12}$$

where $Y_{t-\ell} = \widehat{U}_{t-\ell}^1 = 0$ for all $\ell \geq t$. Setting the initial values to zero when computing the residuals recursively on any iteration is asymptotically negligible (see Lemma 2). Note that the superscript on the parameter matrices refers to the iteration in which those parameters are computed, and the subscript is the usual lag order. We compute the second iteration of the IOLS procedure by plugging $\widehat{U}_t^1$ into (9) yielding $\widehat{X}^1$. Note that $\widehat{X}^1 = (\widehat{X}_{\bar{q}+1}^1, ..., \widehat{X}_T^1)$,

[4] Lemmas 4.1 and 4.2 in Ng and Perron (1995) show that determining the truncation lag proportional to $\ln(T)$ guarantees that the difference between the residuals from the truncated VAR model and the ones obtained from the infinite VAR process is $o_p\left(T^{-1/2}\right)$ uniformly in $\tilde{p}$.

where $\widehat{X}_t^1 = [Y_t - \widehat{U}_t^1, Y_{t-1}, ..., Y_{t-p}, \widehat{U}_{t-1}^1, ..., \widehat{U}_{t-q}^1]'$, is a function of the estimates obtained in the first iteration: $\widehat{\beta}^1$. Similarly as in (11), we obtain $\widehat{\beta}^2$ and its correspondent set of residuals recursively as in (12). The j^{th} iteration of the IOLS estimator is thus given by

$$\widehat{\beta}^j = \left[R' \left(\widehat{X}^{j-1} \widehat{X}^{j-1\prime} \otimes I_K\right) R\right]^{-1} R' \left(\widehat{X}^{j-1} \otimes I_K\right) \text{vec}(Y). \tag{13}$$

We stop the IOLS algorithm when estimates of β converge. We assume that $\widehat{\beta}^j$ converges if $\| \widehat{U}^j - \widehat{U}^{j-1} \| \leq \epsilon$ holds from some exogenously defined criterion ϵ, where $\|.\|$ accounts for the Frobenius norm. If the IOLS fails to converge, we adopt the consistent HR estimator, $\widehat{\beta}^1$, given in (11). The maximum number of iterations is set to 1,000 and $\epsilon = 10^{-5}$. We also notice from our simulations that sample size, number of free parameters, system dimension, and the values of β play an important role on the convergence rates of the IOLS estimator. In general, we find that convergence rates increase monotonically with T (see discussion in Sections 3 and 4).

3 Theoretical Properties

This section provides theoretical results regarding the consistency and the asymptotic distribution of the IOLS estimator. A previous attempt to establish these results have been made by Hannan and McDougall (1988). They prove the consistency of the IOLS estimator considering the univariate ARMA(1,1) specification, but no formal result is provided for the asymptotic normality. Overall, this section differs from the work of Hannan and McDougall (1988) in important ways. First, we derive the consistency and the asymptotic normality for the general weak and strong VARMA(p,q) models; and second, our theory explicitly accounts for the effects of setting initial values equal to zero when updating the residuals on each iteration.

Similarly as in Section 2, we define our baseline weak VARMA(p,q) model expressed in its Echelon form as in (14) and its more compact notation in (15):

$$A_0 Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + ... + A_p Y_{t-p} + A_0 U_t + M_1 U_{t-1} + ... + M_q U_{t-q}, \tag{14}$$

$$Y_t = \left(X_t' \otimes I_K\right) R\beta + U_t, \tag{15}$$

where U_t is a sequence of uncorrelated random variables.

We base our asymptotic results on the general theory for iterative estimators developed by Dominitz and Sherman (2005).[5] Their approach relies on the concept of the Asymptotic Contraction Mapping (ACM). Denote $(\mathbb{B}, d)$ as a metric space where $\mathbb{B}$ is the closed ball centered in β and d is a distance function; $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space, where Ω is a sample space, $\mathcal{A}$ is a σ-field of subsets of Ω and $\mathbb{P}$ is the probability measure on $\mathcal{A}$; and $K_T^{\omega}(.)$ is a function defined on $\mathbb{B}$, with $\omega \in \Omega$. From the definition in Dominitz and Sherman, 2005, p. 841, *"The collection $\{K_T^{\omega}(.) : T \geq 1, \omega \in \Omega\}$ is an ACM on $(\mathbb{B}, d)$ if there exist a constant $c \in [0, 1)$ that does not depend on T or ω, and sets $\{\mathcal{A}_T\}$ with each $\mathcal{A}_T \subseteq \Omega$ and $\mathbb{P}\mathcal{A}_T \to 1$ as $T \to \infty$, such that for each $\omega \in \mathcal{A}_T$, $K_T^{\omega}(.)$ maps $\mathbb{B}$ to itself and for all $x, y \in \mathbb{B}$, $d\left(K_T^{\omega}(x), K_T^{\omega}(y)\right) \leq cd(x, y)$"*. As pointed out by Dominitz and Sherman (2005), if a collection is an ACM, then it will have a unique fixed point in $(\mathbb{B}, d)$, where the fixed point now depends on the sample characteristics, i.e. T and ω. Additionally, their ACM definition nests the case where the population mapping is a fixed deterministic function (Dominitz and Sherman, 2005, p. 840).

Definition 1 (General Mapping) *We define the sample mapping $\widehat{N}_T\left(\widehat{\beta}^j\right)$ and its population counterpart $N\left(\beta^j\right)$ as follows:*

i. $\widehat{\beta}^{j+1} = \widehat{N}_T\left(\widehat{\beta}^j\right) = \left[\frac{1}{T-\bar{q}} \sum_{t=\bar{q}+1}^{T} \widetilde{X}_t^{j\prime} \widetilde{X}_t^j\right]^{-1} \left[\frac{1}{T-\bar{q}} \sum_{t=\bar{q}+1}^{T} \widetilde{X}_t^{j\prime} Y_t\right]$,

ii. $\beta^{j+1} = N\left(\beta^j\right) = \mathbb{E}\left[\widetilde{X}_{\infty,t}^{j\prime} \widetilde{X}_{\infty,t}^j\right]^{-1} \mathbb{E}\left[\widetilde{X}_{\infty,t}^{j\prime} Y_t\right]$,

where $\widetilde{X}_t^j = \left[\left(\widehat{X}_t^{j\prime} \otimes I_K\right) R\right]$ and $\widetilde{X}_{\infty,t}^j = \left[\left(X_{\infty,t}^{j\prime} \otimes I_K\right) R\right]$ have dimensions $(K \times n)$ and denote the regressors computed on the j^{th} iteration; n is the number of free parameters (excluding the distinct covariance parameters) in the model; $\widehat{X}_t^j = \text{vec}(Y_t - \widehat{U}_t^j, Y_{t-1}, ..., Y_{t-p}, \widehat{U}_{t-1}^j, ..., \widehat{U}_{t-q}^j)$, $X_{\infty,t}^j = \text{vec}(Y_t - U_t^j, Y_{t-1}, ..., Y_{t-p}, U_{t-1}^j, ..., U_{t-q}^j)$; $\bar{q} = \max\{p, q\}$; and the $(n \times 1)$ vectors $\widehat{\beta}^j$ and β^j stack the estimates obtained from the sample and population mappings, respectively. Notably, $\widehat{N}_T\left(\widehat{\beta}^j\right)$ and its population counterpart map from $\mathbb{R}^n$ to $\mathbb{R}^n$. The sample and population mappings differ in two important ways. First, the population mapping is a deterministic function, whereas its sample counterpart is stochastic. Second, $\widehat{U}_t^j$ is obtained recursively by

$$\widehat{U}_t^j = Y_t - \left[\widehat{A}_0^j\right]^{-1} \left[\widehat{A}_1^j Y_{t-1} - ... - \widehat{A}_p^j Y_{t-p} - \widehat{M}_1^j \widehat{U}_{t-1}^j - ... - \widehat{M}_q^j \widehat{U}_{t-q}^j\right], \text{ for } t = 1, ..., T, \quad (16)$$

[5] Pastorello et al. (2003) develop a similar general asymptotic theory for iterative estimators that also rests on the contracting property and hence on the unique fixed point condition.

where $Y_{t-\ell} = \widehat{U}^j_{t-\ell} = 0$ for all $\ell \geq t$, while U^j_t is also computed recursively in the same fashion as (16), but assumes the pre-sample values are known, i.e. $Y_{t-\ell}$ and $U^j_{t-\ell}$ are known for all $\ell \geq t$. Note that $N(\beta) = \beta$, which implies that when evaluated on the true vector of parameters, the population mapping maps the vector β to itself. This implies that if the population mapping is an ACM then β is a unique fixed point of $N(\beta)$ (see Dominitz and Sherman, 2005, p. 841).

For theoretical reasons, such that we can formally handle the effect of initial values when computing the residuals recursively, define the infeasible sample mapping as

$$\breve{\beta}^{j+1} = \breve{N}_T\left(\breve{\beta}^j\right) = \left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}^{j\prime}_{\infty,t}\widetilde{X}^{j}_{\infty,t}\right]^{-1}\left[\frac{1}{T-\bar{q}}\sum_{t=\bar{q}+1}^{T}\widetilde{X}^{j\prime}_{\infty,t}Y_t\right]. \tag{17}$$

The infeasible sample mapping in (17) differs from the sample mapping because it is a function of U^j_t rather than $\widehat{U}^j_t$. This is a stochastic version of the population mapping and it is extensively used when deriving the consistency and the asymptotic normality of the IOLS estimator. Lemma 2 shows that, when evaluated at the same vector of estimates, $\widehat{N}_T\left(\widehat{\beta}^j\right)$ converges uniformly to its infeasible counterpart as $T \longrightarrow \infty$, which implies that setting the starting values to zero in (16) does not matter asymptotically.

To formally derive the consistency and asymptotic normality of the IOLS estimator, we impose the following assumptions:

B.1 (Stability, Invertibility, and Uniqueness) Let Y_t be a stable and invertible K-dimensional VARMA(p,q) process. Moreover, assume Y_t is uniquely identified and expressed in Echelon form as in (14) with known Kronecker indices.

B.2 (Disturbances - strong mixing) Let U_t be a $(K \times 1)$ vector of innovations with $K \geq 1$. The disturbances U_t are strictly stationary with $\mathbb{E}(U_t) = 0$, $Var(U_t) = \Sigma_u$, $Cov(U_{t-i}, U_{t-j}) = 0$ for all $i \neq j$ and satisfy the following two conditions:

i. $\mathbb{E}|U_t|^{4+2\nu} < \infty$,

ii. $\sum_{\kappa=0}^{\infty}\{\alpha_u(\kappa)\}^{\nu/(2+\nu)} < \infty$, for some $\nu > 0$,

where $\alpha_u(l) = \sup_{\substack{\mathcal{D}\in\sigma(U_i, i\geq t+l)\\ \mathcal{C}\in\sigma(U_i, i\leq t),}}|Pr(\mathcal{C}\bigcap\mathcal{D}) - Pr(\mathcal{C})Pr(\mathcal{D})|$ are strong mixing coefficients of order $l \geq 1$, with $\sigma(U_i, i \leq t)$ and $\sigma(U_i, i \geq t+l)$ being the σ-fields generated by $\{U_i : i \leq t\}$ and $\{U_i : i \geq t+l\}$, respectively.

B.3 (Contraction and Stochastic Equicontinuity) Define the $(n \times n)$ infeasible sample gradient as $\breve{V}_T\left(\breve{\beta}^j\right) = \frac{\partial \breve{N}_T\left(\breve{\beta}^j\right)}{\partial \breve{\beta}^{j\prime}}$ and its population counterpart as $V\left(\beta^j\right) = \frac{\partial N\left(\beta^j\right)}{\partial \beta^{j\prime}}$; and $\mathbb{B}$ as the closed ball centered at β satisfying invertibility and stability conditions in Assumption B.1. Assume that the following hold:

i. The maximum eigenvalue associated with $V(\beta) = \left.\frac{\partial N\left(\beta^j\right)}{\partial \beta^{j\prime}}\right|_{\beta}$ is smaller than one in absolute value.

ii. $\sup_{\phi \in \mathbb{B}} \|\breve{V}_T(\phi)\| = O_p(1)$, with $\phi \in \mathbb{B}$.

Assumption B.1 provides the general regularity conditions governing the VARMA(p,q) model. Assumption B.2 establishes the mixing conditions which satisfy the weak VARMA definition. Notably, these mixing conditions are valid for a wide range of nonlinear models that allow weak VARMA representations (Francq and Zakoian (1998), Francq et al. (2005) and Francq and Zakoian (2005)). Item i. in Assumption B.3 suffices to guarantee that the IOLS mapping is an ACM on $(\mathbb{B}, E_n)$. Lemma 1 provides the sample counterpart of this result, making it possible to verify on every iteration whether the sample mapping is an ACM. Albeit the result in Lemma 1 is computationally easy to obtain, we could not pin down the eigenvalues of the population counterpart of Lemma 1 solely as function of the parameters matrices eigenvalues. Numerical simulation indicates that the maximum eigenvalue of $V(\beta)$ depends on the elements of the parameters matrices in (14) rather than on their eigenvalues. Furthermore, we note that the population mapping is not an ACM if some of the eigenvalues of the AR and MA components have opposite signs and are close to one in absolute value.

Define $Z = \left[R'\left(H \otimes I_K\right)R\right]^{-1}$, $H = \text{plim}\left[\frac{1}{T}\sum_{t=\bar{q}+1}^{T} X_t X_t'\right]$, $J = \left[I_n - V(\beta)\right]^{-1}$, and $\mathcal{I} = \sum_{\ell=-\infty}^{\infty} \mathbb{E}\left\{\left[R'\left(X_t \otimes I_K\right)U_t\right]\left[R'\left(X_{t-\ell} \otimes I_K\right)U_{t-\ell}\right]'\right\}$. Theorem 1 gives the consistency and asymptotic distribution of the IOLS estimator for the general weak VARMA(p,q) model.

Theorem 1 *Suppose Assumptions B.1, B.2, and B.3 hold. Then,*

i. $\left|\widehat{\beta} - \beta\right| = o_p(1)$ *as* $j,\ T \longrightarrow \infty$;

ii. $\sqrt{T}\left[\widehat{\beta} - \beta\right] \xrightarrow{d} \mathcal{N}\left(0, JZ\mathcal{I}Z'J'\right)$ *as* $j,\ T \longrightarrow \infty$ *and* $\frac{\ln(T)}{j} = o(1)$.

Proof. See Appendix.

The proof proceeds as follows. Item i. in Theorem 1 requires that both $N(\phi)$ and $\widehat{N}_T(\phi)$ are ACMs on $(\mathbb{B}, E_n)$, where E_n is the Euclidean metric on $\mathbb{R}^n$ and $\mathbb{B}$ is a closed ball centered

at β. Lemma 5 gives that $\widehat{N}_T(\phi)$ is an ACM on $(\mathbb{B}, E_n)$ and hence $\left|\widehat{N}_T(\phi) - \widehat{N}_T(\gamma)\right| \leq \kappa|\phi - \gamma|$ holds, with $\gamma, \phi \in \mathbb{B}$ and $\kappa \in [0,1)$. Additionally, the sample mapping has an unique fixed point in $\mathbb{B}$ given by $\widehat{\beta}$, and Lemma 5 establishes that $\widehat{\beta}^j$ converges uniformly in $\mathbb{B}$ to $\widehat{\beta}$. It follows that consistency of the IOLS estimator is obtained by using the standard fixed point theorem and Lemma 3 (the sample mapping converges uniformly in probability to its population counterpart). Next, the limiting distribution of $\widehat{\beta}$ also demands $N(\phi)$ and $\widehat{N}_T(\phi)$ to be ACMs on $(\mathbb{B}, E_n)$. Furthermore, it requires uniform convergence of $\breve{N}_T(\phi)$ to $\widehat{N}_T(\phi)$ in $\phi \in \mathbb{B}$ (Lemma 2), $\sup_{\phi,\gamma\in\mathbb{B}} \left\| \left[\breve{\Lambda}_T(\phi,\gamma) - \Lambda(\phi,\gamma)\right](\phi - \gamma)\right\| = o_p(1)$ (Lemma 4), and $\sqrt{T}\left|\widehat{\beta}^j - \widehat{\beta}\right| = o_p(1)$ as $j,\ T \longrightarrow \infty$ and $\frac{\ln(T)}{j} = o(1)$ (Lemma 6). It follows that the asymptotic normality result simplifies to the limiting behaviour of $[I_n - V(\beta)]^{-1}\sqrt{T}\left[\breve{N}_T(\beta) - \beta\right]$. We follow the work of Francq and Zakoian (1998) and Dufour and Pelletier (2014), and use the central limit theorem of Ibragimov (1962) that encompasses strong mixing processes such as the one in Assumption B.2. In turn, $\sqrt{T}\left[\breve{N}_T(\beta) - \beta\right] \xrightarrow{d} \mathcal{N}(0, Z\mathcal{I}Z')$ as $j,\ T \longrightarrow \infty$. This yields an asymptotic variance that is a function of $\mathcal{I}$ rather than the usual $R'(H \otimes \Sigma_u)R$ term that appears when U_t is an *i.i.d.* process (see Corollary 1).

Lemma 1 and $\widehat{\beta}$ can be used to compute the empirical counterparts of $V(\beta)$, H, Z and Σ_u, yielding a feasible estimate of the asymptotic variance. Specifically, $\mathcal{I}$ can be consistently estimated by the Newey-West covariance estimator,

$$\widehat{\mathcal{I}} = \frac{1}{T - \bar{q}} \sum_{\ell=-m_T}^{m_T} \left[1 - \frac{|\ell|}{m_T + 1}\right] \sum_{t=\bar{q}+1+|\ell|}^{T} \left\{ \left[R'\left(\widehat{X}_t \otimes I_K\right)\widehat{U}_t\right] \times \left[R'\left(\widehat{X}_{t-\ell} \otimes I_K\right)\widehat{U}_{t-\ell}\right]' \right\}, \tag{18}$$

where $m_T^4/T \to 0$ with $T, m_T \to \infty$.

The practical implication of the violation of item i. in Assumption B.3 is that the IOLS estimator does not converge even for large T. If this is the case, the asymptotic results in this section cannot be implemented. By contrast, we note from our simulations that if the IOLS estimator converges, the contraction property assumption is satisfied, i.e. the modulus of the largest eigenvalue of $\widehat{V}_T\left(\widehat{\beta}\right)$ is strictly less than unity. It is also possible to have a DGP satisfying item i. in Assumption B.3 and the IOLS estimator does not converge. This follows because Lemma 5 holds only asymptotically, and convergence in finite sample requires that both the population and sample mappings are ACMs, with the latter holding

on every iteration. In turn, it is possible to use Lemma 1 and evaluate the modulus of the largest eigenvalue of $\widehat{V}_T\left(\widehat{\beta}^j\right)$ on every iteration to verify whether the IOLS estimator is in its region of convergence. The Monte Carlo study shows that when the population mapping is an ACM (item i. in Assumption B.3 holds), convergence rates increase monotonically with T.[6] Additionally, the Monte Carlo section discusses small sample adjustments that improve the convergence rates of the IOLS estimator.

Finally, we derive the consistency and limiting distribution of the IOLS estimator under the stronger assumption that the true data generation process is a strong VARMA process, i.e. U_t is an *i.i.d.* process. The proof works in a similar fashion as the one in Theorem 1. Specifically, all the auxiliary results (Lemmas 2, 3, 4, 5, and 6) hold under the *i.i.d.* assumption; and the limiting distribution of $[I_n - V(\beta)]^{-1}\sqrt{T}\left[\breve{N}_T(\beta) - \beta\right]$ is now derived using the central limit theorem for *m.d.s.*

Corollary 1 *Suppose Assumptions B.1 and B.3 hold. Additionally, assume*

i. U_t *is an i.i.d. process with* $\mathbb{E}(U_t) = 0$, $Var(U_t) = \Sigma_u$ *and finite fourth moment.*

Then, item i. in Theorem 1 holds and $\sqrt{T}\left[\widehat{\beta} - \beta\right] \xrightarrow{d} \mathcal{N}(0, JZR'(H \otimes \Sigma_u)RZ'J')$ *as* $j, T \longrightarrow \infty$ *and* $\frac{\ln(T)}{j} = o(1)$.

Proof. See Appendix.

4 Monte Carlo Study

This section provides results on the finite sample performance of VARMA models estimated with the IOLS methodology. We compare the IOLS estimator with estimators possessing very different asymptotic and computational characteristics. We report results considering five methods: the MLE, the two-stage method (HR) of Hannan and Rissanen (1982), the three-step procedure (HK) of Hannan and Kavalieris (1984), the two-stage method (DJ2) of Dufour and Jouini (2014), and the multivariate version of the three-step procedure (KP) of Koreisha and Pukkila (1990) as discussed in Koreisha and Pukkila (2004) and Kascha (2012). To broadly analyse and assess the performance of the IOLS estimator, we design simulations covering different sample sizes (from 50 to 1,000 observations), system sizes

[6] Figure S.1 in the online Supplement displays finite sample convergence rates (heat-map) of the IOLS estimator for ARMA(1,1) models defined as $y_t = \beta_1 y_{t-1} + u_t + \beta_2 u_{t-1}$. We show that convergence rates improve dramatically when sample size increases from $T = 100$ to $T = 10,000$ in the entire set of parameters satisfying item i. in Assumption B.3.

($K = 3$, $K = 10$, $K = 20$, $K = 40$, and $K = 52$), Kronecker indices, dependencies among the variables, and both weak and strong processes.[7]

We simulate stable, invertible, and unique VARMA(1,1) models as

$$A_0 Y_t = A_1 Y_{t-1} + A_0 U_t + M_1 U_{t-1}. \tag{19}$$

Uniqueness is imposed through the Echelon form transformation, which implies $A_0 = M_0$ in (19) and requires a choice of Kronecker indices. We discuss results considering six DGPs. DGPs I and II set all Kronecker indices to one, which implies that $A_0 = I_K$ and A_1 and M_1 are full matrices. These DGPs differ with respect to the eigenvalues assigned to the parameter matrices. The eigenvalues of A_1 and M_1 in DGP I are constant and equal to 0.5, whereas the eigenvalues in DGP II take positive, negative and near-to-zero values. Precisely, the eigenvalues of A_1 and M_1 are $(0.80, 0.20, 0.05)'$ and $(0.90, 0.02, 0.20)'$, respectively.[8] For DGPs III, IV, V, and VI, the first k Kronecker indices are set to one, while the remaining $K - k$ Kronecker indices are set to zero, so that $\mathbf{p} = (p_1, p_2, ..., p_K)'$ with $p_i = 1$ for $i \leq k$ and $p_i = 0$ for all $i > k$. Specifically, DGP III has $k = 1$, while DGPs IV, V and VI have $k = 2$, $k = 3$, and $k = 6$, respectively. The free parameters in DGPs III, IV, V, and VI are based on real data and chosen as the estimates obtained by fitting VARMA(1,1) models to the datasets in Section 5. DGPs III-VI are particularly relevant because they reduce dramatically the number of free parameters in (19), while yielding rich dynamics in the MA component of the standard representation of (19).[9]

Weak VARMA(1,1) models are obtained by generating U_t as in Romano and Thombs, 1996, p. 591, with $U_t = \prod_{\ell=0}^{m} \varepsilon_{t-\ell}$, where $m = 3$ and ε_t is a zero mean *i.i.d.* process with covariance matrix I_K. This procedure yields uncorrelated innovations satisfying the mixing conditions stated in Assumption B.2. We summarize results for each specification using two measures: MRRMSE and Share. MRRMSE accounts for the mean of the relative root median squared error (RRMSE) measures of all parameters, where RRMSE is the ratio of the root median squared error (RMSE) obtained from a given estimator over the RMSE of the HR estimator. RRMSE measures lower than one indicate that the HR estimator is outperformed by the alternative estimator. Share is the frequency a given estimator returns

[7] For sake of brevity, results for the strong VARMA specifications are in the online Supplement.

[8] File "DGPsfile.csv" (available online) contains the true values used in all simulations in this section.

[9] Because DGPs III-VI imply that A_0 is an invertible lower triangular matrix, multiplying (19) by A_0^{-1} yields its standard representation, $Y_t = A_1^* Y_{t-1} + U_t + M_1^* U_{t-1}$, where $A_1^* := A_0^{-1} M_1 = (A_{1,K\times k}^*, 0)$ and $M_1^* := A_0^{-1} M_1$ is a full matrix.

the lowest RRMSE over all the free parameters.[10] The MRRMSE and Share measures are only computed using the replications that achieved convergence and satisfy Assumption B.1.[11] We discard the initial 500 observations and fix the number of replications to 5,000, unless otherwise stated. We report only a fraction of the entire set the Monte Carlo results. The complete set of tables is in Section S.3 in the online Supplement.

A valid concern that emerges when estimating DGPs I and II with the IOLS estimator is its rather low convergence rates for small sample sizes. Nevertheless, convergence rates of the IOLS estimator increase monotonically with T in both DGPs and system dimensions ($K = 3$ and $K = 10$), which leave us with the conclusion that this is a small sample issue. This is in line with our theoretical results, as Lemma 1 holds when evaluated at the true vector of parameters and Lemma 5 holds only asymptotically. We address the causes for the IOLS estimator failing to achieve convergence and identify three reasons for this finite sample anomaly. The IOLS fails when, first, it generates at some iteration a non-invertible VARMA model that contains no roots on the unit circle; second, it generates a non-stable VARMA model with no roots on the unit circle; third, it converges to multiple (finite) fixed points. We address these issues individually in the IOLS estimator. Namely, first, in the case of a non-invertible VARMA model, we convert the non-invertible moving average polynomial to its corresponding invertible representation using Lippi and Reichlin's 1994 procedure and continue iterating. Second, in the instance of a non-stable VARMA model, we factorize the VAR polynomial and replace the eigenvalues greater than one in absolute value by 0.99 and continue iterating. Finally, in the case of convergence to multiple fixed points, we choose the one which minimizes the determinant of the covariance matrix of the residuals and stop iterating.[12] Implementing these small sample adjustments improves the convergence rates and, most importantly, does not come at expense of worse results.[13] In short, considering DGPs I and II with sample sizes of $T \leq 400$, convergence rates increase an average of 34% and 56% for $K = 3$ and $K = 10$, respectively, while MRRMSE measures increase by only 3% and 1% for $K = 3$ and $K = 10$, respectively. Numbers are even more

[10] As an example, when $K = 3$ and $\mathbf{p} = (1, 0, 0)'$, $n = 6$, i.e. there are six free parameters (excluding the distinct covariance parameters) to be estimated. If the IOLS estimator has a Share of 67%, it implies that the IOLS estimator delivers the lowest RRMSE in four out of those six free parameters. This measure is particularly informative when dealing with systems with large number of free parameters.

[11] Throughout this section, we assume an estimator converges if its final estimates satisfy Assumption B.1. For the IOLS and MLE estimators, convergence also implies numerical convergence of their respective algorithms.

[12] Section S.3.2 details the implementation of the small sample adjustments and discusses their finite sample performance.

[13] See Tables S.6-S.13 in the online Supplement.

favourable when considering DGP II, as convergence rates increase by 59% and 89% for $K = 3$ and $K = 10$, respectively, while MRRMSE measures increase by only 5% and 2% for $K = 3$ and $K = 10$, respectively. Despite the improvements in the convergence rates, we take a conservative stand and report only results for the IOLS estimator computed without any finite sample adjustments (Tables 1 and 2). When appropriate, we further discuss the convergence rates and performances of the small sample adjustments.

The first set of Monte Carlo simulations addresses the finite sample performance of the IOLS estimator in small sized ($K = 3$) VARMA(1,1) models. We simulate weak VARMA(1,1) models considering DGPs I, II and III. DGPs I and II yield 18 free parameters in the A_0, A_1, and M_1 parameter matrices, while DGP III has 6 free parameters in these matrices. We consider samples of 50, 100, 150, 200, and 400 observations. Table 1 presents the results. First, we find that the MLE estimator is dominant in DGP I. This is not surprising, because MLE is known to perform well on specifications where the absolute eigenvalues are bounded away from zero and one (Kascha (2012)). The DGP II specification is numerically more difficult to handle, yielding lower rates of convergence for both the IOLS and MLE estimators. Nevertheless, the IOLS estimator is the one which delivers the best results considering both MRRMSE and Share measures. This indicates that if convergence is achieved, the IOLS estimator is able to handle systems with near-to-zero eigenvalues more efficiently than the benchmark MLE estimator. Furthermore, when considering the IOLS computed with the small sample adjustments discussed before, convergence rates for the DGP II increase to 53%, 67%, 72%, 75%, and 82% for $T = 50$, $T = 100$, $T = 150$, $T = 200$, and $T = 400$, respectively. The IOLS estimator is very competitive in DGP III and presents the highest Share measure in all sample sizes. Simulating DGPs I, II, and III for $K = 10$ provides a comparable picture (see detailed discussion in the online Supplement).

Table 2 displays results covering the finite sample performance of the IOLS estimator in medium and large sized systems. We simulate DGPs III, IV, V, and VI for $K \in (10, 20, 40, 52)$, which mimic our empirical study. To the best of our knowledge, this is the first study to consider such high-dimensional VARMA models in a Monte Carlo study. The sample size and number of replications are set to $T = 400$ and $1,000$, respectively. These are high-dimensional models with the number of free parameters (excluding the distinct covariance parameters), n, varying from 20 to 624. We find that the IOLS estimator presents the best relative performance (in terms of both the MRRMSE and Share measures) for

large systems ($K = 40$ and $K = 52$). These findings hold for all DGPs. The relative performance of the HK and IOLS estimators are outstanding, with an average improvement with respect to the HR estimator of up to 65%. On average, the IOLS outperforms the HK estimator in 15% (in terms of the MRRMSE measure). Considering systems with $K = 10$ and $K = 20$, the HK estimator is the one that delivers the best performance for most specifications, while the IOLS estimator is constantly ranked as the second best estimator.[14] Therefore, the IOLS estimator considerably improves its performance when estimating high-dimensional restricted models with $A_0 \neq I_K$, while remaining a feasible alternative (average convergence rate of 95%). These results motivate the use of the Echelon form transformation in the fashion of DGPs III-VI and the IOLS estimator as feasible alternative to high-dimensional VARMA(1,1) models.

Overall, we conclude that the IOLS estimator is a competitive alternative and compares favourable with its competitors in a variety of cases: small sample sizes, small sized systems with near-to-zero eigenvalues, and large sized systems with many Kronecker indices set to zero. The MLE and HK estimators also present remarkable performances in terms of RMSE, which is in line with previous studies (Kascha (2012)).

5 Empirical Application

In this section, we analyse the competitiveness of VARMA models estimated with the IOLS procedure to forecast macroeconomic variables. We forecast three key macroeconomic variables: industrial production (IPS10), interest rate (FYFF), and CPI inflation (PUNEW). We assess VARMA forecast performance under different system dimensions and forecast horizons.

5.1 Data and Setup

We use US monthly seasonally adjusted data from the Stock and Watson (2006) dataset, which runs from 1959:1 through 2003:12. The choice of using seasonally adjusted data is in line with the large datasets literature (see Stock and Watson (2002a,b), De Mol et al. (2008), and Bańbura et al. (2010)), as seasonal data intensifies the "curse of dimensionality" by requiring the inclusion of extra lags. Therefore, as usual in this literature, a practitioner

[14] The online Supplement presents results where IOLS outperforms the HK estimator for $K = 10$ and $K = 20$ in DGPs IV-VI with $T = 200$.

should be cautious when using the framework discussed in this section for seasonal data. As in Carriero et al. (2011), we use 52 macroeconomic variables that represent the main categories of economic indicators. We work with five system dimensions: $K \in (3, 10, 20, 40, 52)$. We construct four different datasets (one to four) for each system size where $10 \leq K \leq 40$, as a way to assess robustness of the VARMA framework when dealing with different explanatory variables. When selecting the variables, we try to keep a balance among the three main categories of data: real economy, money and prices, and financial market.[15] The series are transformed, as in Carriero et al. (2011), in such a way that they are approximately stationary. The forecasting exercise is performed in pseudo real time, with a fixed rolling window of 400 observations. All models considered in the exercise are estimated in every window. We perform 115 out-of-sample forecasts considering six different horizons: one- (Hor:1), two- (Hor:2), three- (Hor:3), six- (Hor:6), nine- (Hor:9) and twelve- (Hor:12) steps-ahead.

We compare the different VARMA specifications with five alternative methods: AR(1), ARMA(1,1), VAR(p^*), BVAR, and factor models.[16] ARMA(1,1) models are estimated using the IOLS and maximum-likelihood estimators. We estimate the BVAR model with the normal-inverted Wishart prior as in Bańbura et al. (2010), so that the prior is adjusted to accommodate the fact that our variables are approximately stationary.[17] We report results considering three BVAR specifications. The first specification, BVAR_{SC}, is obtained by setting the hyperparameter φ (tightness parameter) to the value which minimizes the SC criterion over a grid of $\varphi \in (2.0e-5, 0.0005, 0.002, 0.008, 0.018, 0.072, 0.2, 1, 500)$.[18] The second specification, $\text{BVAR}_{0.2}$, sets $\varphi = 0.2$, which is the default choice in the package Regression Analysis of Time Series (RATS) and it is the benchmark model in Carriero et al. (2011). The third specification, BVAR_{opt}, follows from Bańbura et al. (2010) and chooses $\varphi \in (2.0e-5, 0.0005, 0.002, 0.008, 0.018, 0.072, 0.2, 1, 500)$ which minimizes the in-sample one-step-ahead root mean squared forecast error in the last 24 months of the sample. For the three different specifications, we choose the lag length that minimizes the SC criterion.[19] We grid search over φ and the optimal lag length on every rolling window.

[15] Table S.14 in the online Supplement reports the details of the different datasets.

[16] The lag length p^* in the VAR(p^*) specification is obtained by minimizing the AIC criterion.

[17] See the online Supplement (Section S.4.1) for an extended discussion on the BVAR framework implemented in this section.

[18] This grid follows Carriero et al. (2011) and it is broad enough to include both very tight ($\varphi = 2.0e-5$) and loose ($\varphi = 500$) hyperparameters.

[19] The maximum lag length is set to be 15, 8, and 6 for $K \leq 10$, $20 \leq K \leq 40$, and $K = 52$, respectively.

Factor models summarize a large number of predictors in only a few number of factors. We adopt the two-step procedure as in Stock and Watson (2002a,b). In the first step, factors ($\{F_t\}_{t=1}^T$) are extracted via principal components, whereas the second step consists of projecting $y_{i,t+h}$ onto $\left(\widehat{F}_t', ..., \widehat{F}_{t-\ell}', y_{i,t-1}, ..., y_{i,t-r}\right)$, with $\ell \geq 0$ and $r > 0$. We determine the lag orders by minimizing the Schwarz (SC) criterion, and choose the number of factors according to the SC and IC_{p3} criteria, denoted as FM_{SC} and FM_{IC3}, respectively (Stock and Watson (2002a)). Factors are computed using only the variables available on the respective dataset.

When dealing with empirical data, the Kronecker indices which are required to express a VARMA representation in its Echelon form are unknown. We determine them using three strategies. The first one imposes rank reduction on the parameter matrices of the VARMA(1,1) model in a similar fashion as the DGPs III, IV, V, and VI in the Monte Carlo section. This follows Carriero et al. (2011), who show that reduced rank VAR (RRVAR) models perform well when forecasting macroeconomic variables using large datasets. By specifying the Kronecker indices as $\mathbf{p} = (p_1, p_1, ..., p_K)'$ with $p_i = 1$ for $i \leq k$ and $p_i = 0$ for all $i > k$, the rank of both parameter matrices of the standard VARMA representation $A_1^* = A_0^{-1}A_1$ and $M_1^* = A_0^{-1}M_1$ reduces to k, in that the VAR(∞) representation of the VARMA(1,1) model is a RRVAR(∞) of rank k. We report results for $k \in (1, 2, 3, 4, 5, 6)$ denoted as $\mathbf{p}_{k=1}$, $\mathbf{p}_{k=2}$, $\mathbf{p}_{k=3}$, $\mathbf{p}_{k=4}$, $\mathbf{p}_{k=5}$, and $\mathbf{p}_{k=6}$. Notably, the restrictions and number of free parameters in $\mathbf{p}_{k=1}$, $\mathbf{p}_{k=2}$, $\mathbf{p}_{k=3}$, and $\mathbf{p}_{k=6}$ are analogous to DGPs III, IV, V, and VI, respectively. The second and third strategies estimate the Kronecker indices directly from the data. Specifically, the second strategy adopts the Hannan-Kavalieris algorithm (denoted as $\mathbf{p}_{HK}$), which consists of choosing Kronecker indices that minimize the SC_K criterion (Lütkepohl, 2007, p. 503).[20] We implement this procedure on every rolling window for $K = 3$ and $K = 10$, and on the first rolling window for medium and large sized datasets ($K = 20$, $K = 40$, and $K = 52$). We then carry the estimated Kronecker indices to the subsequent rolling windows. The maximum value of the Kronecker indices is set to one.

The third strategy is based on Poskitt (1992) and reviewed in Lütkepohl and Poskitt (1996). We refer to this method as the Lütkepohl-Poskitt procedure and denote it as $\mathbf{p}_{LP}$.[21] The Lütkepohl-Poskitt procedure is based on the property of the Echelon form that the restrictions on the k^{th} equation do not depend on the Kronecker indices $p_i > p_k$. The

[20] See Section S.4.2 in the online Supplement for a complete description of the Hannan-Kavalieris procedure.
[21] We thank an anonymous referee for suggesting this procedure.

Lütkepohl-Poskitt procedure consists of estimating each of the K equations in the system by least squares (using the residuals from VAR($\tilde{p}$) instead of the true disturbances) and evaluating model selection criteria for each of the K equations separately.[22] Because the Kronecker indices are estimated equation by equation, the Lütkepohl-Poskitt searches for the first local minimum of the criterion for each equation, and hence it suits particularly well high-dimensional models. The Lütkepohl-Poskitt procedure is consistent under suitable conditions (Poskitt (1992)). Specifically, if Assumptions B.1 and B.3 hold and U_t is an *i.i.d.* process, choosing $\tilde{p}$ with the AIC criterion and setting the model selection criterion as the SC criterion meet these conditions. The Lütkepohl-Poskitt procedure has two important advantages compared to the Hannan-Kavalieris algorithm. First, it does not impose a pre-specified upper bound to the Kronecker indices. This implies that very general and highly complex VARMA(p,q) models with $p, q \geq 1$ and up to K nonzero Kronecker indices can be selected. Despite its general approach, the Lütkepohl-Poskitt procedure also attains some degree of sparseness and parsimony, as it penalizes for the number of freely varying parameters in each equation. Second, the Lütkepohl-Poskitt procedure is preferable from a computational point of view, as it allows the estimation of the Kronecker indices on every rolling window for all system sizes. To assess how these different strategies fit the data, we report two SC criteria, denoted as SC_K and SC_3. SC_K is the standard SC criterion computed with the entire $(K \times K)$ covariance matrix of the residuals, whereas SC_3 uses only the (3×3) upper block of the residuals covariance matrix, as this contains the covariance matrix of the three key macroeconomic variables. The SC_3 criterion, therefore, is a measure of fit that is solely related to the variables that we are ultimately interested in. All VARMA specifications are estimated with the IOLS estimator discussed in Section 2.[23]

We compare different models using the out-of-sample relative mean squared forecast error (RelMSFE) computed using the AR(1) as a benchmark. We choose this benchmark, because it makes easy the comparison across the different datasets and system dimensions. We assess the predictive accuracy with the Diebold and Mariano (1995) test. The use of this test is justified given our focus on forecasts obtained through rolling windows.

[22] See Section S.4.3 in the online Supplement for a complete description of the Lütkepohl-Poskitt procedure.

[23] If the IOLS does not converge, we implement the small sample adjustments discussed in Section 4. If convergence is not achieved, we adopt the consistent initial estimates: the two-stage HR estimator.

5.2 Results

We organize the results as follows. Table 3 reports a summary of the forecast results across the different datasets and presents, for each system size, three panels. The first panel reports the frequency (in percentage points) for which at least one of the VARMA specifications ($\mathbf{p}_{k=1}$, $\mathbf{p}_{k=2}$, $\mathbf{p}_{k=3}$, $\mathbf{p}_{k=4}$, $\mathbf{p}_{k=5}$, $\mathbf{p}_{k=6}$, $\mathbf{p}_{HK}$, and $\mathbf{p}_{LP}$) outperforms (delivers the lowest RelMSFE measures) the assigned group of competitors in a given forecast horizon.[24] Similarly, the second and third panel display the frequencies for which the most restricted, $\mathbf{p}_{k=1}$, and general, $\mathbf{p}_{LP}$, specifications outperform the assigned group of competitors. We consider five groups of competitors: AR, ARMA, VAR, FM, and BVAR. The AR group collects the AR(1) specification; ARMA has the ARMA(1,1) specification estimated with IOLS and MLE; VAR contains the VAR(p^*) model; FM gathers the different factor model specifications, namely the FM_{SC} and FM_{IC3}; and the BVAR collects the three BVAR specifications: BVAR_{SC}, $\text{BVAR}_{0.2}$ and BVAR_{opt}. Finally, Table 4 compares the performance of the IOLS estimator with the DJ2, HK, and KP estimators. We report the frequency (in percentage points) for which the IOLS estimator outperforms the alternative estimators for each forecast horizon. A comprehensive set of results is available in Section S.4.4 in the online Supplement (Tables S.15 to S.28), making it possible to assess the forecast performance of all VARMA specifications and the alternative model competitors in all datasets and system sizes up to the level of the key macroeconomic variables. Convergence rates for the IOLS estimator are also reported for all VARMA specifications. In what follows, we draw from these tables when discussing the empirical results.

Starting with $K = 3$, we find that VARMA models largely outperform the AR, VAR, FM, and the BVAR groups up to the sixth-step-ahead forecast (see Table 3). Specifically, the $\mathbf{p}_{k=1}$ and $\mathbf{p}_{k=2}$ specifications are the ones which deliver the best results. VARMA models also outperform the ARMA(1,1) specifications, which reinforces the idea that modelling the three key macroeconomic variables in a multivariate context pays off. Taking into account all forecast horizons, VARMA models deliver the best forecast in 67% of the cases.

We now summarize the results for the medium and large datasets, ($K = 10$, $K = 20$, $K = 40$, and $K = 52$). We start by discussing the results for $K = 10$. For short horizons (Hor:1 - Hor:6), VARMA models deliver the lowest RelMSFE in 63% of the cases. Com-

[24] Percentages are computed across the three key macroeconomic variables and the four datasets discussed in Section 5.1.

pared with the BVAR group, VARMA models remain dominant, delivering more accurate forecasts in 81% of the cases (Hor:1 - Hor:6). Moreover, the $\mathbf{p}_{k=1}$ and $\mathbf{p}_{k=2}$ specifications are the ones that usually deliver the lowest RelMSFE measures among the VARMA specifications. Specifically, $\mathbf{p}_{k=1}$ minimizes the SC_3 criterion in all datasets, which indicates that choosing the Kronecker indices that minimize the SC_3 criterion pays off in terms of forecast accuracy. Increasing the number of nonzero Kronecker indices from three to six, $\mathbf{p}_{k=3}$, $\mathbf{p}_{k=4}$, $\mathbf{p}_{k=5}$, and $\mathbf{p}_{k=6}$ specifications, delivers stable RelMSFE measures which are less often the best among the competitors. This is in line with the large datasets literature, which documents that imposing restrictions on the parameter matrices of a standard VAR model improves forecast accuracy (De Mol et al. (2008), Carriero et al. (2011), and Bańbura et al. (2010)). Choosing the Kronecker indices according to the general Lütkepohl-Poskitt procedure and the Hannan-Kavalieris algorithm typically yields more complex models (up to eight nonzero Kronecker indices), stable RelMSFE measures, and a slightly less accurate forecast performance. Specifically, while the $\mathbf{p}_{LP}$ generally outperforms the AR, ARMA, and VAR for short horizons, it is outperformed by FM, and BVAR specifications. With regard to the estimation of VARMA models, the IOLS estimator works well, presenting an average convergence rate of 93%. Additionally, its relative performance with respect to the DJ2, HK, and KP estimators is positive. Considering the $\mathbf{p}_{k=1}$ and the $\mathbf{p}_{k=2}$ specifications, the IOLS estimator outperforms its linear competitors in 81% of the cases (Table 4).

We now discuss the results for $K = 20$. Overall, VARMA models are very competitive in the short horizons (Hor:1 - Hor:6), outperforming the AR, ARMA, VAR, FM and BVAR groups in 90%, 71%, 94%, 79%, and 88% of the cases, respectively. Results are also stable across the different datasets and VARMA specifications, showing robustness of the VARMA framework. Differently from the $K = 10$ scenario, there is not a clear winner among the VARMA specifications, although $\mathbf{p}_{k=1}$ remains very competitive and minimizes the SC_3 criterion. While setting the number of nonzero Kronecker indices as $k \in (2, 3, 4, 5, 6)$ typically does not improve forecast accuracy ($\mathbf{p}_{k=6}$ in Dataset 3 is the exception), the data driven Lütkepohl-Poskitt procedure emerges as a strong competitor and yields the most accurate forecasts among all VARMA specifications in 29% of the cases. The IOLS estimator presents rates of convergence averaging 94% (across all specifications) and usually delivers more accurate forecasts than the DJ2, HK, and KP estimators in the long horizons. The strong performance of the HK estimator in the short horizons is in line with the Monte

Carlo results.

Considering the case of large datasets ($K = 40$), we find that VARMA models stay very competitive, outperforming the AR, ARMA, VAR, FM, and BVAR groups in 92%, 81%, 96%, 58%, and 82% of the cases, respectively. The general $\mathbf{p}_{LP}$ specification now emerges as the clear winner among the VARMA specifications, as it produces the most accurate forecasts in 46% of the cases. In turn, the Lütkepohl-Poskitt procedure is able to select the relevant Kronecker indices to forecast the key macroeconomic variables and preserves some degree of sparseness in the parameter matrices, which is important when forecasting using rich datasets. The good performance of the $\mathbf{p}_{LP}$ specification is also due to the use of the IOLS estimator, as the IOLS estimator outperforms the DJ2, HK, and KP estimators in 68%, 71%, and 94% of the cases (Table 4). Indeed, the Monte Carlo simulations show that the IOLS estimator delivers an outstanding performance in large sized systems, ($K = 40$ and $K = 52$). We report marginal gains in terms of forecast accuracy when moving from the $\mathbf{p}_{k=1}$ specification to the more general models where the number of nonzero Kronecker indices are $k \in (2, 3, 4, 5, 6)$. Similarly to the case of the $\mathbf{p}_{LP}$ specification, the IOLS estimator systematically delivers more accurate forecasts than the alternative estimators. Considering the specifications with $k \in (1, 2, 3, 4, 5, 6)$, the IOLS estimator outperforms the DJ2, HK, and KP estimators in 79%, 87%, and 86% of the cases, respectively (Table 4), and improves the RelMSFE measures in 19%, 33%, and 40% on average for the DJ2, HK, and KP estimators, respectively (Hor:1 - Hor:6).[25] Finally, the IOLS estimator remains a robust alternative, achieving convergence in 93% of the rolling windows (all specifications).

We now turn our attention to the results using the entire dataset ($K = 52$). When comparing the performance of VARMA models with the BVAR group, the former delivers lower RelMSFE measures in 72% of the cases. This is a strong result in favour of VARMA models, since BVAR specifications are known to be very competitive when forecasting with large datasets. Overall, factor models deliver the best performance. Among the VARMA specifications, $\mathbf{p}_{k=1}$ and $\mathbf{p}_{k=2}$ deliver the best results. Increasing the number of nonzero Kronecker indices to $k \in (3, 4, 5, 6)$ delivers at most marginal improvements in terms of RelMSFE measures. When the IOLS estimator presents low convergence rates ($\mathbf{p}_{LP}$ and $\mathbf{p}_{HK}$ specifications), VARMA models become less competitive. Alternatively, using the RelMSFE measures obtained with the DJ2, HK, and KP estimators does not help either, as

[25] Results available upon request.

they yield no qualitative improvement in terms of forecast hierarchy when compared to the IOLS based measures. On the comparison of the IOLS and the alternative linear estimators, the IOLS estimator largely outperforms (presents values greater than 50%) the DJ2, HK, and KP estimators for the $\mathbf{p}_{k=1}$, $\mathbf{p}_{k=2}$, $\mathbf{p}_{k=3}$, $\mathbf{p}_{k=4}$, $\mathbf{p}_{k=5}$, and $\mathbf{p}_{k=6}$ specifications. Specifically, the IOLS estimator outperforms the DJ2, HK, and KP estimators in 83%, 98%, and 94% of the cases, respectively (Table 4), and improves the RelMSFE measures compared to the DJ2, HK, and KP estimators an average of 33%, 54%, and 70%, respectively, (Hor:1 - Hor:3).[26] Finally, convergence rates for these specifications are 100%.

To sum up the results of this section, VARMA models estimated using the IOLS estimator are generally very competitive and able to beat the four most prominent competitors in this type of study: AR, ARMA, factor models and BVAR models. This finding is especially present for the $\mathbf{p}_{k=1}$ and $\mathbf{p}_{LP}$ specifications. VARMA results are also stable across the different datasets and Kronecker indices specifications, indicating that the framework adopted is fairly robust. Considering all system sizes and datasets, VARMA specifications deliver the lowest RelMSFE measures for the one-, two-, three-, and six-month-ahead forecast in 63% of the cases, indicating that VARMA models are indeed strong candidates to forecast key macroeconomic variables using small, medium and large sized datasets. It is particularly relevant to highlight the performance of VARMA models relative to the BVAR models, as VARMA models systematically deliver more accurate forecasts for all datasets. The IOLS estimator is a valid alternative to deal with large and complex VARMA systems (convergence rates averaging 92%) and compares favourably with the main linear competitors (more accurate forecasts in 69% of the cases). Finally, our findings reinforce two important aspects: using the Echelon form transformation either in the fashion of DGPs III-VI or using the Lütkepohl-Poskitt procedure is a powerful tool to deal with high-dimensional models; and that IOLS estimator is particularly suitable to estimate high-dimensional VARMA models, as the Monte Carlo simulations and empirical results suggest.

6 Conclusion

This paper addresses the issue of modelling and forecasting key macroeconomic variables using rich (small, medium and large sized) datasets. We propose the use of VARMA models as a feasible framework for this task. We overcome the natural difficulties in estimating

[26] Results available upon request.

medium- and high-dimensional VARMA models with the MLE framework by adopting the IOLS estimator.

We establish the consistency and asymptotic distribution for the IOLS estimator by considering the general weak and strong VARMA(p,q) models. It is also important to point out that our theoretical results are obtained under weak assumptions that are compatible with the quasi-maximum-likelihood (QMLE) estimator. The extensive Monte Carlo study shows that the IOLS estimator is feasible and consistent in small and high-dimensional systems. Furthermore, the IOLS estimator outperforms the MLE and other linear estimators, in terms of mean squared error, in a variety of scenarios: when T is small; disturbances are weak; near-to-zero eigenvalues; and high-dimensional models ($K = 40$ and $K = 52$). The empirical results show that VARMA models perform better than AR(1), ARMA(1,1), VAR, BVAR, and factor models for different system sizes and datasets. We find that VARMA models estimated with the IOLS estimator are very competitive at forecasting short horizons (one-, two-, three- and six-month-ahead horizons) in small, medium, and large sized datasets. In particular, the $\mathbf{p}_{k=1}$ and the general $\mathbf{p}_{LP}$ specifications most often emerge as the ones which produce the most accurate results among all VARMA specifications. Finally, we find that VARMA models estimated with the IOLS estimator usually deliver better forecasts than the ones estimated with the alternative linear estimators.

7 Appendix

Proof of Theorem 1: Denote $\phi = (\phi_1, \phi_2, ..., \phi_n)'$ with $\phi \in \mathbb{B}$, as an n-dimensional vector collecting the parameter estimates of a general weak VARMA(p,q) model. We start proving the consistency of the IOLS estimator. This proof follows analogous steps as in Theorems 2 and 4 in Dominitz and Sherman (2005) (DS henceforth). The steps in our proof are related to the primitive conditions in these theorems and hence we refer to them to easy the exposition. From DS, if $N(\phi)$ is an ACM on $(\mathbb{B}, E_n)$, then $N(\phi)$ is also a contraction mapping. Item i. in Assumption B.3 implies that $N\left(\beta^j\right)$ is an ACM, so that $\left|N\left(\beta^{j-1}\right) - N(\beta)\right| \leq \kappa \left|\beta^{j-1} - \beta\right|$ and $N(\beta) = \beta$ hold. From Lemma 5, the sample mapping is an ACM on $(\mathbb{B}, E_n)$ with a fixed point $\widehat{\beta}$ in the closed set $\mathbb{B}$, $\widehat{\beta} \in \mathbb{B}$ (condition ii. in DS's Theorem 4). First, bound $\left|\widehat{\beta} - \beta\right|$ as

$$\left|\widehat{\beta} - \beta\right| \leq \left|\beta^j - \beta\right| + \left|\widehat{\beta} - \beta^j\right|. \tag{20}$$

To show that $\left|\beta^j - \beta\right|$ converges to zero, rewrite it as

$$\left|\beta^j - \beta\right| = \left|N\left(\beta^{j-1}\right) - N\left(\beta\right)\right| \leq \kappa \left|\beta^{j-1} - \beta\right|. \tag{21}$$

Recursive substitution of (21) yields $\left|\beta^j - \beta\right| \leq \kappa^j \left|\beta^0 - \beta\right|$. As $j \longrightarrow \infty$, $\left|\beta^j - \beta\right| = o\left(1\right)$, and hence the first term on the right-hand side of (20) converges to zero. It remains to show that $\left|\widehat{\beta} - \beta^j\right|$ has order $o_p\left(1\right)$. We bound $\left|\widehat{\beta} - \beta^j\right|$ using the auxiliary result given by (S.46) in Lemma 5, so that

$$\left|\widehat{\beta} - \beta^j\right| \leq \kappa^j \left|\beta^0 - \widehat{\beta}\right| + \left(\sum_{i=0}^{j-1}\right) \kappa^i \left[\sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T\left(\phi\right) - N\left(\phi\right)\right|\right], \tag{22}$$

where β^0 is the vector collecting the initial estimates for the population mapping. As $j \longrightarrow \infty$, with $\kappa \in (0, 1]$, $\beta^0 \in \mathbb{B}$, $\widehat{\beta} \in \mathbb{B}$, and $\mathbb{B}$ is a closed ball centered at β, it follows that (22) reduces to

$$\left|\widehat{\beta} - \beta^j\right| \leq \sup_{\phi \in \mathbb{B}} \left|\widehat{N}_T\left(\phi\right) - N\left(\phi\right)\right| \left[\frac{1}{1-\kappa}\right]. \tag{23}$$

Because the second term in brackets on the right-hand side of (23) is bounded and the first term has order $o_p\left(1\right)$, (Lemma 3), $\left|\widehat{\beta} - \beta^j\right| = o_p(1)$, which implies $\left|\widehat{\beta} - \beta\right| = o_p(1)$ and completes the proof of the consistency result. The consistency result is equivalent to condition i. in DS's Theorem 4, as $\widehat{\beta}^j$ converges uniformly in $\mathbb{B}$ to $\widehat{\beta}$ (Lemma 5).

We now turn our attention to the asymptotic distribution of the IOLS estimator. First, write

$$\sqrt{T}\left|\widehat{\beta}^j - \beta\right| \leq \sqrt{T}\left|\widehat{\beta} - \beta\right| + \sqrt{T}\left|\widehat{\beta}^j - \widehat{\beta}\right|. \tag{24}$$

Lemma 6 gives that the second term on the right-hand side of (24) is $o_p\left(1\right)$ as $j, T \longrightarrow \infty$ and $\frac{\ln(T)}{j} = o(1)$ (condition iii. in DS's Theorem 4). Next, rewrite $\sqrt{T}\left[\widehat{\beta} - \beta\right]$ as

$$\sqrt{T}\left[\widehat{\beta} - \beta\right] = \sqrt{T}\left[\widehat{N}_T\left(\beta\right) - \breve{N}_T\left(\beta\right)\right] + \sqrt{T}\left[\widehat{N}_T\left(\widehat{\beta}\right) - \widehat{N}_T\left(\beta\right)\right] + \sqrt{T}\left[\breve{N}_T\left(\beta\right) - \beta\right]. \tag{25}$$

Lemma 2 gives $\sqrt{T}\left[\widehat{N}_T\left(\beta\right) - \breve{N}_T\left(\beta\right)\right] = O_p\left(T^{-1/2}\right)$. Using the mean value theorem,

rewrite the second term on the right-hand side of (25) as

$$\sqrt{T}\left[\widehat{N}_T\left(\widehat{\beta}\right)-\widehat{N}_T\left(\beta\right)\right]=\sqrt{T}\Big\{\left[\widehat{N}_T\left(\widehat{\beta}\right)-\breve{N}_T\left(\widehat{\beta}\right)\right]+\left[\breve{N}_T\left(\beta\right)-\widehat{N}_T\left(\beta\right)\right]\Big\}+ \sqrt{T}\Big\{\breve{\Lambda}_T\left(\widehat{\beta},\beta\right)\left[\widehat{\beta}-\beta\right]\Big\}, \tag{26}$$

where $\breve{\Lambda}_T\left(\widehat{\beta},\beta\right)=\int_0^1 \breve{V}_T\left(\widehat{\beta}+\xi\left(\widehat{\beta}-\beta\right)\right)d\xi$. Because the first term has order $O_p(T^{-1/2})$ and $\breve{\Lambda}_T\left(\widehat{\beta},\beta\right)\left[\widehat{\beta}-\beta\right]$ converges uniformly to its population counterpart (Lemma 4), (25) reads

$$\sqrt{T}\left[\widehat{\beta}-\beta\right]=\sqrt{T}\left[\left[I_n-\Lambda\left(\widehat{\beta},\beta\right)\right]^{-1}\left[\breve{N}_T\left(\beta\right)-\beta\right]\right]. \tag{27}$$

Notably, Lemma 4 fulfills the condition v. in DS's Theorem 4. From item (i.) in Theorem 1, $\widehat{\beta}$ converges in probability to β as $j\longrightarrow\infty$ with $T\longrightarrow\infty$, implying that $\Lambda\left(\widehat{\beta},\beta\right)$ converges in probability to $V\left(\beta\right)$. It follows that (27) reduces to

$$\sqrt{T}\left[\widehat{\beta}-\beta\right]=\sqrt{T}\left[\left[I_n-V\left(\beta\right)\right]^{-1}\left[\breve{N}_T\left(\beta\right)-\beta\right]\right]. \tag{28}$$

The next step consists of proving condition iv. in DS's Theorem 4. As $\breve{N}_T\left(\beta\right)$ is evaluated at the true vector of parameters and $T\longrightarrow\infty$, it follows that $\sqrt{T}\left[\breve{N}_T\left(\beta\right)-\beta\right]$ reads

$$\sqrt{T}\left[\breve{N}_T\left(\beta\right)-\beta\right]=\left[R'\left[\left(\frac{1}{T}\sum_{t=\bar{q}+1}^{T}X_tX_t'\right)\otimes I_K\right]R\right]^{-1}\times \left[\frac{1}{\sqrt{T}}\sum_{t=\bar{q}+1}^{T}R'\left(X_t\otimes I_K\right)U_t\right]. \tag{29}$$

Recall that the last term is not an *m.d.s*, because U_t is not an *i.i.d.* process. Assumption B.1 allows a VMA(∞) representation of Y_t of the form $Y_t=\sum_{i=0}^{\infty}\Theta_iU_{t-i}$ with $\Theta_0=I_K$. As discussed in Francq and Zakoian (1998), a stationary process which is a function of a finite number of current and lagged values of U_t satisfies a mixing property of the form Assumption B.2. Using the VMA(∞) representation of Y_t, partition X_t in (29) into $X_t=X_t^r+X_t^{r+}$,

such that

$$X_t = \begin{pmatrix} Y_t - U_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p} \\ U_{t-1} \\ \vdots \\ U_{t-q} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{r} \Theta_i U_{t-i} \\ \sum_{i=0}^{r} \Theta_i U_{t-1-i} \\ \vdots \\ \sum_{i=0}^{r} \Theta_i U_{t-p-i} \\ U_{t-1} \\ \vdots \\ U_{t-q} \end{pmatrix} + \begin{pmatrix} \sum_{i=r+1}^{\infty} \Theta_i U_{t-i} \\ \sum_{i=r+1}^{\infty} \Theta_i U_{t-1-i} \\ \vdots \\ \sum_{i=r+1}^{\infty} \Theta_i U_{t-p-i} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = X_t^r + X_t^{r+}. \tag{30}$$

Rewrite $\frac{1}{\sqrt{T}} \sum_{t=\bar{q}+1}^{T} R' \left(X_t \otimes I_K \right) U_t$ as

$$\frac{1}{\sqrt{T}} \sum_{t=\bar{q}+1}^{T} R' \left(X_t \otimes I_K \right) U_t = \frac{1}{\sqrt{T}} \sum_{t=\bar{q}+1}^{T} R' \left(X_t^r \otimes I_K \right) U_t + \frac{1}{\sqrt{T}} \sum_{t=\bar{q}+1}^{T} R' \left(X_t^{r+} \otimes I_K \right) U_t. \tag{31}$$

Auxiliary results in Dufour and Pelletier (2014, Theorem 4.2) show that the second term on the right-hand side of (31) converges uniformly to zero in T as $r \to \infty$. It follows that first term on the right-hand side of (31) satisfies the strong mixing conditions of the form Assumption B.2. We are now in position to use Ibragimov's 1962 central limit theorem for strong mixing processes (see also Dufour and Pelletier (2014, Lemma A.2)). This yields $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} R' \left(X_t^r \otimes I_K \right) U_t \xrightarrow{d} \mathcal{N} \left(0, \mathcal{I}_r \right)$. From Francq and Zakoian, 1998, p. 157, $\mathcal{I}_r \xrightarrow{p} \mathcal{I}$ as $r \to \infty$, such that as $T, r \to \infty$

$$\sqrt{T} \left[\breve{N}_T \left(\beta \right) - \beta \right] \xrightarrow{d} \mathcal{N} \left(0, Z \mathcal{I} Z' \right), \tag{32}$$

where $\mathcal{I} = \sum_{\ell=-\infty}^{\infty} \mathbb{E} \left\{ \left[R' \left(X_t \otimes I_K \right) U_t \right] \left[R' \left(X_{t-\ell} \otimes I_K \right) U_{t-\ell} \right]' \right\}$, and $Z = \left[R' \left(H \otimes I_K \right) R \right]^{-1}$ with $H = \operatorname{plim} \frac{1}{T} \sum_{t=\bar{q}+1}^{T} X_t X_t'$. The final result of this Theorem is obtained by combining the first element of the right-hand side of (28) with (32), such that

$$\sqrt{T} \left[\widehat{\beta} - \beta \right] \xrightarrow{d} \mathcal{N} \left(0, J Z \mathcal{I} Z' J' \right), \tag{33}$$

where $J = \left[I_n - V \left(\beta \right) \right]^{-1}$.

Proof of Corollary 1: The consistency result follows directly from item i. in Theorem 1.

To show item ii., rewrite $\sqrt{T}\left[\widehat{\beta}-\beta\right]$ using the same arguments as in the proof of Theorem 1, such that

$$\sqrt{T}\left[\widehat{\beta}-\beta\right]=\sqrt{T}\left[\left[I_n-V(\beta)\right]^{-1}\left[\breve{N}_T(\beta)-\beta\right]\right]. \tag{34}$$

The limiting distribution of (34) depends on the limiting behaviour of,

$$\sqrt{T}\left[\breve{N}_T(\beta)-\beta\right]=\left[R'\left[\left(\frac{1}{T}\sum_{t=\bar{q}+1}^{T}X_tX_t'\right)\otimes I_K\right]R\right]^{-1}\left[\frac{1}{\sqrt{T}}\sum_{t=\bar{q}+1}^{T}R'\left(X_t\otimes I_K\right)U_t\right]. \tag{35}$$

Because U_t is an *i.i.d.* process, it follows that $R'\left(X_t\otimes I_K\right)U_t$ is an *m.d.s*, in that the central limit theorem for *m.d.s* can be used to show that (35) converges in distribution to

$$\sqrt{T}\left[\breve{N}_T(\beta)-\beta\right]\xrightarrow{d}\mathcal{N}\left(0,ZR'\left(H\otimes\Sigma_u\right)RZ'\right), \tag{36}$$

where $Z=\left[R'\left(H\otimes I_K\right)R\right]^{-1}$ and $H=\operatorname{plim}\frac{1}{T}\sum_{t=\bar{q}+1}^{T}X_tX_t'$. Define $J=\left[I_n-V(\beta)\right]^{-1}$ and combine (34) with (36), and the asymptotic distribution of the IOLS estimator for the strong VARMA(p,q) models reads,

$$\sqrt{T}\left[\widehat{\beta}-\beta\right]\xrightarrow{d}\mathcal{N}\left(0,JZR'\left(H\otimes\Sigma_u\right)RZ'J'\right). \tag{37}$$

Lemma 1 *Assume Assumptions B.1 and B.2 hold, then* $\widehat{V}_T\left(\widehat{\beta}^j\right)=\frac{\partial\widehat{N}_T\left(\widehat{\beta}^j\right)}{\partial\widehat{\beta}^{j\prime}}$ *is given by:*

$$\begin{aligned}\widehat{V}_T\left(\widehat{\beta}^j\right)=\Bigg\{&\left[I_1\otimes W^{j^{-1}}\right]\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\Bigg\{\left(Y_t'\otimes I_n\right)\left(I_K\otimes R'\right)\times\\&\left[\left(I_1\otimes\mathbb{K}_{K,f}\otimes I_K\right)\left(I_f\otimes vec\left(I_K\right)\right)\right]\frac{\partial vec\left(\widehat{X}_t^j\right)}{\partial\widehat{\beta}^{j\prime}}\Bigg\}+\\&\Bigg\{\left[\left(\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\tilde{X}_t^{j\prime}Y_t\right)'\otimes I_n\right]\left[-\left(W^j\right)^{-1}\otimes\left(W^j\right)^{-1}\right]\times\\&\left[\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\Bigg\{\left[\left(I_{n^2}+\mathbb{K}_{n,n}\right)\left(I_n\otimes\tilde{X}_t^{j\prime}\right)\left(R'\otimes I_K\right)\times\right.\\&\left.\left[\left(I_f\otimes\mathbb{K}_{K,1}\otimes I_K\right)\left(I_f\otimes vec\left(I_K\right)\right)\right]\frac{\partial vec\left(\widehat{X}_t^{j\prime}\right)}{\partial\widehat{\beta}^{j\prime}}\Bigg\}\right]\Bigg\}\end{aligned} \tag{38}$$

where $\mathbb{K}$ *is the commutation matrix,* $f=K(p+q+1)$, $\bar{q}=\max\{p,q\}$,

$W^j = \left(\frac{1}{T-\bar{q}}\sum_{t=1+\bar{q}}^{T}\tilde{X}_t^{j\prime}\tilde{X}_t^j\right)$, *and* $\frac{\partial vec(\widehat{X}_t^j)}{\partial\widehat{\beta}^{j\prime}} = \frac{\partial vec(\widehat{X}_t^{j\prime})}{\partial\widehat{\beta}^{j\prime}}$ *with*

$$\frac{\partial vec\left(\widehat{X}_t^j\right)}{\partial\widehat{\beta}^{j\prime}} = vec\left(-\frac{\partial\widehat{U}_t^j}{\partial\widehat{\beta}^{j\prime}}, 0_{K,n}, \cdots 0_{K,n}, \frac{\partial\widehat{U}_{t-1}^j}{\partial\widehat{\beta}^{j\prime}}, \cdots, \frac{\partial\widehat{U}_{t-q}^j}{\partial\widehat{\beta}^{j\prime}}\right) \quad and \tag{39}$$

$$\begin{aligned}\frac{\partial\widehat{U}_t^j}{\partial\widehat{\beta}^{j\prime}} = &\left\{\left(\widehat{A}_0^{j^{-1}}\left[\widehat{A}_1^j Y_{t-1}+...+\widehat{A}_p^j Y_{t-p}+\widehat{M}_1^j\widehat{U}_{t-1}^j+...+\widehat{M}_q^j\widehat{U}_{t-q}^j\right]\right)'\otimes\widehat{A}_0^{j^{-1}}\right\}\times\\ &\left[I_{K^2}:0:...:0\right]R-\left[\left(Y_{t-1}',...,Y_{t-p}',\widehat{U}_{t-1}^{j\prime},...,\widehat{U}_{t-q}^{j\prime}\right)\otimes\widehat{A}_0^{j^{-1}}\right]\left[0:I_{K^2(p+q)}\right]R-\\ &\widehat{A}_0^{j^{-1}}\left[M_1^j\frac{\partial\widehat{U}_{t-1}^j}{\partial\widehat{\beta}^{j\prime}}+...+M_q^j\frac{\partial\widehat{U}_{t-q}^j}{\partial\widehat{\beta}^{j\prime}}\right].\end{aligned} \tag{40}$$

Proof. See Section S.1 in the online Supplement.

Lemma 2 *Assume Assumptions B.1 and B.2 hold. Then,*

$\sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi)-\breve{N}_T(\phi)\right\| = O_p\left(T^{-1}\right)$.

Proof. See Section S.1 in the online Supplement.

Lemma 3 *Assume Assumptions B.1, B.2, and B.3 hold. Then,*

$\sup_{\phi\in\mathbb{B}}\left\|\widehat{N}_T(\phi)-N(\phi)\right\| = o_p(1)$ *as* $T\longrightarrow\infty$.

Proof. See Section S.1 in the online Supplement.

Lemma 4 *Assume Assumptions B.1, B.2, and B.3 hold. Then,*

$\sup_{\phi,\gamma\in\mathbb{B}}\left\|\left[\widehat{\Lambda}_T(\phi,\gamma)-\Lambda(\phi,\gamma)\right](\phi-\gamma)\right\| = o_p(1)$ *as* $T\longrightarrow\infty$,

$\sup_{\phi,\gamma\in\mathbb{B}}\left\|\left[\breve{\Lambda}_T(\phi,\gamma)-\Lambda(\phi,\gamma)\right](\phi-\gamma)\right\| = o_p(1)$ *as* $T\longrightarrow\infty$.

Proof. See Section S.1 in the online Supplement.

Lemma 5 *Assume Assumptions B.1, B.2, and B.3 hold, then,* $\widehat{N}_T(\phi)$ *is an ACM on* $(\mathbb{B},E_n)$*, with* $\phi\in\mathbb{B}$ *and it has fixed point denoted by* $\widehat{\beta}$*, such that* $\left|\widehat{\beta}^j-\widehat{\beta}\right| = o_p(1)$ *uniformly in* $\mathbb{B}$ *as* $j,T\longrightarrow\infty$.

Proof. See Section S.1 in the online Supplement.

Lemma 6 *Assume Assumptions B.1, B.2, and B.3 hold. If*

i. $\widehat{N}_T(\phi)$ *is an ACM on* $(\mathbb{B},E_n)$

then, $\sqrt{T}\left|\widehat{\beta}^j-\widehat{\beta}\right| = o_p(1)$ *as* $j,T\longrightarrow\infty$ *and* $\frac{\ln(T)}{j} = o(1)$.

Proof. See Section S.1 in the online Supplement.

Acknowledgments: This paper previously circulated under the title "Forecasting Medium and Large Datasets with Vector Autoregressive Moving Average (VARMA) Models". We are grateful to the Editor (Oliver Linton), an associate editor, and two anonymous referees for their valuable comments and suggestions. We are also indebted to Emmanuel Guerre, Cristina Scherrer, and the seminar participants at Queen Mary (Econometrics Reading Group), the 14th Brazilian Time Series and Econometrics School (2011), the Econometric Society Australasian Meeting (2011), the Royal Economic Society Annual Conference (2011) and the 4th International Conference on Computational and Financial Econometrics (2010). Gustavo Fruet Dias acknowledges support from CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation. The usual disclaimer applies.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 667–674.

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. In L. D. Mehar, R. (Ed.), *System Identification*, pp. 27–96. New York: Academic Press.

Athanasopoulos, G., D. S. Poskitt, and F. Vahid (2012). Two canonical varma forms: Scalar component models vis-à-vis the echelon form. *Econometric Reviews 31*(1), 60–83.

Athanasopoulos, G. and F. Vahid (2008). A complete varma modelling methodology based on scalar components. *Journal of Time Series Analysis 29*, 533–554.

Bańbura, M., D. Giannone, and L. Reichlin (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics 25*, 71–92.

Bernanke, B., J. Boivin, and P. S. Eliasz (2005, January). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics 120*(1), 387–422.

Blanchard, O. and D. Quah (1989). The dynamic effects of aggregate demand and supply disturbances. *American Economic Review 79*, 655–673.

Carriero, A., G. Kapetanios, and M. Marcellino (2011). Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics 26*, 735–761.

Chan, J. C., E. Eisenstat, and G. Koop (2016). Large bayesian varmas. *Journal of Econometrics 192*(2), 374 – 390. Innovations in Multiple Time Series Analysis.

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: Is bayesian regression a valid alternative to principal components? *Journal of Econometrics 146*, 318–328.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*, 253–263.

Dominitz, J. and R. Sherman (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory 21*, 838–863.

Dufour, J.-M. and T. Jouini (2005). Asymptotic distribution of a simple linear estimator for varma models in echelon form. In P. Duchesne and B. Rémillard (Eds.), *Statistical Modeling and Analysis for Complex Data Problems*, pp. 209–240. Springer US.

Dufour, J.-M. and T. Jouini (2014). Asymptotic distributions for quasi-efficient estimators in echelon varma models. *Computational Statistics & Data Analysis 73*, 69–86.

Dufour, J.-M. and D. Pelletier (2014). Practical methods for modelling weak varma processes: Identification, estimation and specification with a macroeconomic application. Manuscript.

Dufour, J.-M. and D. Stevanović (2013). Factor-augmented varma models with macroeconomic applications. *Journal of Business & Economic Statistics 4*, 491–506.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalised dynamic factor model: Identification and estimation. *Review of Econometrics and Statistics 82*(4), 540–554.

Francq, C., R. Roy, and J.-M. Zakoian (2005). Diagnostic checking in arma models with uncorrelated errors. *Journal of the American Statistical Association 100*, 532–544.

Francq, C. and J.-M. Zakoian (1998). Estimating linear representations of nonlinear processes. *Journal of Statistical Planning and Inference 68*, 145–165.

Francq, C. and J.-M. Zakoian (2005). Recent results for linear time series models with non independent innovations. In P. Duchesne and B. RMillard (Eds.), *Statistical Modeling and Analysis for Complex Data Problems*, pp. 241–265. Springer US.

Gourieroux, C. and A. Monfort (2015). Revisiting identification and estimation in structural varma models. Technical report, CREST.

Hannan, E. J. (1969). The identification of vector mixed autoregressive-moving average systems. *Biometrika 56*(1), 223–225.

Hannan, E. J. (1976). The identification and parameterization of armax and state space forms. *Econometrica 44*(4), 713–723.

Hannan, E. J. and M. Deistler (1988). *The Statistical Theory of Linear Systems*. New York: John Wiley and Sons.

Hannan, E. J. and L. Kavalieris (1984, Sep). Multivariate linear time series models. *Advances in Applied Probability 16*(3), 492–561.

Hannan, E. J. and A. J. McDougall (1988). Regression procedures for arma estimation. *Journal of the American Statistical Association 83*, 490–498.

Hannan, E. J. and J. Rissanen (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika 69*, 81–94.

Ibragimov, A. (1962). Some limit theorems for stationary processes. *Theory of Probability and its Applications 7*, 349–382.

Kapetanios, G. (2003, June). A note on an iterative least squares estimation method for arma and varma models. *Economics Letters 79*(3), 305–312.

Kascha, C. (2012). A comparison of estimation methods for vector autoregressive moving-average models. *Econometric Reviews 31*, 297–324.

Koreisha, S. and T. Pukkila (1990). A generalized least squares approach for estimation of autoregressive moving-average models. *Journal of Time Series Analysis 11*, 139–151.

Koreisha, S. and T. Pukkila (2004). The specification of vector autoregressive moving average models. *Journal of Statistical Computation and Simulation 74*, 547–565.

Lippi, M. and L. Reichlin (1994). Var analysis, nonfundamental representations, blaschke matrices. *Journal of Econometrics 63*(1), 307 – 325.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer- Verlag.

Lütkepohl, H. and D. S. Poskitt (1996). Specification of echelon-form varma models. *Journal of Business & Economic Statistics 14*, 69–79.

Ng, S. and P. Perron (1995). Unit root tests in arma models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association 90*, 268–281.

Pastorello, S., V. Patilea, and E. Renault (2003). Iterative and recursive estimation in structural nonadaptive models. *Journal of Business and Economics Statistics 21*(460), 449–509.

Poskitt, D. S. (1992). Identification of echelon canonical forms for vector linear processes using least squares. *The Annals of Statistics 20*(1), 195–215.

Romano, J. P. and L. A. Thombs (1996). Inference for autocorrelations under weak assumptions. *Journal of the American Statistical Association 91*, 590–600.

Rubio-Ramírez, J. F., D. F. Waggoner, and T. Zha (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies 77*, 665–696.

Spliid, H. (1983). A fast estimation method for the vector autoregressive moving average model with exogenous variables. *Journal of the American Statistical Association 78*, 843–849.

Stock, J. H. and M. W. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Stock, J. H. and M. W. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economics Statistics 20*, 147–162.

Stock, J. H. and M. W. Watson (2006). Forecasting with many predictors. Volume 1 of *Handbook of Economic Forecasting*, Chapter 10, pp. 515 – 554. Elsevier.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), pp. 267–288.

Uhlig, H. (2005). What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics 52*(2), 381 – 419.

Table 1: Monte Carlo - Weak VARMA(1,1) models: Small Sized Systems, $K = 3$.

	DGP I						DGP II						DGP III					
	T=50, $n = 18$						T=50, $n = 18$						T=50, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	0.89	**0.81**	1.09	0.95	0.91	1.00	**0.95**	0.97	1.10	0.97	1.08	1.00	0.66	0.77	0.78	0.98	**0.61**
Share (%)	0%	**39%**	39%	0%	17%	6%	0%	**44%**	0%	0%	33%	22%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	97%	47%	73%	84%	85%	45%	98%	25%	83%	72%	91%	24%	99%	66%	96%	88%	95%	72%
	T=100, $n = 18$						T=100, $n = 18$						T=100, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	0.98	**0.91**	1.05	0.99	0.97	1.00	1.00	**0.99**	1.11	1.01	1.13	1.00	0.79	0.85	0.82	1.11	**0.75**
Share (%)	0%	**39%**	17%	6%	11%	28%	11%	**44%**	6%	0%	28%	11%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	99%	67%	88%	95%	94%	86%	99%	37%	93%	78%	97%	59%	99%	70%	99%	91%	100%	90%
	T=150, $n = 18$						T=150, $n = 18$						T=150, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	1.03	1.05	1.06	**1.00**	1.00	1.00	**0.96**	1.02	1.04	1.01	1.09	1.00	0.86	1.29	0.87	1.17	**0.82**
Share (%)	6%	**28%**	22%	6%	11%	**28%**	0%	**56%**	33%	6%	0%	6%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	100%	76%	81%	98%	97%	94%	100%	42%	83%	83%	99%	70%	100%	72%	98%	93%	100%	95%
	T=200, $n = 18$						T=200, $n = 18$						T=200, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	1.04	1.04	1.06	**0.99**	1.01	1.00	**0.94**	1.02	1.02	1.01	1.05	1.00	0.88	1.29	0.88	1.16	**0.85**
Share (%)	11%	**33%**	6%	6%	11%	**33%**	0%	**44%**	33%	11%	0%	11%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	100%	80%	86%	99%	98%	95%	100%	48%	86%	87%	99%	74%	100%	71%	100%	93%	100%	97%
	T=400, $n = 18$						T=400, $n = 18$						T=400, $n = 6$					
	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE	HR	IOLS	DJ2	HK	KP	MLE
MRRMSE	1.00	1.05	1.04	1.03	**1.00**	1.00	1.00	**0.88**	1.07	0.95	1.01	0.92	1.00	**0.90**	1.39	0.92	1.16	0.91
Share (%)	6%	28%	6%	6%	17%	**39%**	0%	**44%**	33%	0%	0%	22%	0%	**67%**	0%	0%	0%	33%
Convergence(%)	100%	88%	95%	100%	99%	97%	100%	57%	88%	91%	99%	80%	100%	71%	100%	95%	100%	99%

We report results for weak VARMA(1,1) models simulated with different Kronecker indices. The first set of results reports results for DGP I, while the second and third set of results display results for weak VARMA(1,1) models simulated from DGPs II and III, respectively. Recall that DGP I and II set all the Kronecker indices to one, $\mathbf{p} = (1, 1, 1)'$, while DGP III sets $\mathbf{p} = (1, 0, 0)'$. DGPs I and II differ with respect to the eigenvalues driving the AR and MA parameter matrices. DGP I has all the eigenvalues associated with both the AR and MA parameter matrices set to 0.5, while DGP II has eigenvalues associated with the AR and MA components given by $(0.80, 0.20, 0.05)'$ and $(0.90, -0.02, -0.20)'$, respectively. The true vector of parameters in DGP III collects the estimates obtained by fitting a VARMA(1,1) model to the first rolling window of a dataset comprising only the three key macroeconomic variables studied in Section 5. File "DGPsfile.csv" (available online) contains the true values. n accounts for the number of free parameters in A_0, A_1, and M_1 parameter matrices. MRRMSE is the mean of the RRMSE measures of all parameters. The lowest MRRMSE is highlighted in bold. RRMSE measures are computed as the ratio of the RMSE (root median squared error) measures obtained from a given estimator over the HR estimator. Share is the percentage over the total number of free parameters for which a given estimator delivers the lowest MRRMSE. The highest Share is highlighted in bold. Convergence is the percentage of replications in which the algorithms converged and yielded invertible and stable models. HR is the two-stage estimator of Hannan and Rissanen (1982); DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage estimator of Hannan and Kavalieris (1984); KP is the multivariate version of the three-step estimator of Koreisha and Pukkila (1990) as formulated in Kascha (2012); and MLE accounts for the maximum-likelihood estimator. The number of replications is set to 5,000.

Table 2: Monte Carlo - Weak VARMA(1,1) models: Medium and Large Sized Systems, $K = 10$, $K = 20$, $K = 40$ and $K = 52$, with $T = 400$.

	DGP III					DGP IV					DGP V					DGP VI				
	K=10, $n = 20$					K=10, $n = 40$					K=10, $n = 60$					K=10, $n = 120$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	0.99	2.00	**0.76**	1.17	1.00	0.94	1.57	**0.89**	1.22	1.00	0.97	1.46	**0.94**	1.20	1.00	**0.98**	1.18	0.99	1.10
Share (%)	0%	25%	5%	**70%**	0%	10%	23%	8%	**60%**	0%	17%	18%	17%	**48%**	0%	10%	32%	**38%**	15%	5%
Convergence(%)	100%	98%	100%	100%	98%	100%	92%	99%	100%	95%	100%	83%	99%	100%	95%	100%	83%	100%	100%	85%
	K=20, $n = 40$					K=20, $n = 80$					K=20, $n = 120$					K=20, $n = 240$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	1.01	2.23	**0.79**	1.49	1.00	0.98	1.76	**0.87**	1.41	1.00	0.98	1.65	**0.88**	1.45	1.00	0.96	1.30	**0.92**	1.38
Share (%)	30%	15%	0%	**55%**	0%	19%	19%	1%	**61%**	0%	13%	14%	7%	**65%**	2%	4%	24%	22%	**47%**	3%
Convergence(%)	100%	100%	100%	100%	82%	100%	97%	100%	100%	80%	100%	94%	100%	100%	75%	100%	77%	100%	100%	81%
	K=40, $n = 80$					K=40, $n = 160$					K=40, $n = 240$					K=40, $n = 480$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	**0.35**	1.30	0.46	1.07	1.00	**0.44**	1.30	0.59	1.28	1.00	**0.47**	1.36	0.62	1.37	1.00	**0.53**	1.35	0.70	1.38
Share (%)	4%	**56%**	0%	40%	0%	9%	**75%**	0%	16%	0%	10%	**83%**	0%	8%	0%	14%	**86%**	0%	1%	0%
Convergence(%)	100%	100%	87%	100%	65%	100%	99%	87%	100%	62%	100%	100%	80%	99%	45%	100%	99%	77%	99%	28%
	K=52, $n = 104$					K=52, $n = 208$					K=52, $n = 312$					K=52, $n = 624$				
	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP	HR	IOLS	DJ2	HK	KP
MRRMSE	1.00	**0.41**	2.01	0.45	2.39	1.00	**0.43**	1.71	0.53	2.16	1.00	**0.48**	1.92	0.58	2.76	1.00	**0.57**	1.63	0.70	2.38
Share (%)	4%	**51%**	0%	45%	0%	4%	**67%**	0%	29%	0%	6%	**82%**	0%	12%	0%	13%	**85%**	1%	1%	0%
Convergence(%)	100%	100%	81%	99%	36%	100%	100%	75%	97%	22%	100%	100%	63%	96%	8%	100%	99%	43%	95%	4%

We report results for weak VARMA(1,1) models simulated with different Kronecker indices and system sizes. The first set of results reports results for DGP III, while the second, third, and four set of results display results for weak VARMA(1,1) models simulated from DGPs IV, V, and VI, respectively. Recall that DGPs III, IV, V, and VI set the first k Kronecker indices to one and the remaining $K - k$ Kronecker indices to zero, so that $\mathbf{p} = (p_1, p_2, ..., p_K)'$ with $p_i = 1$ for $i \leq k$ and $p_i = 0$ for all $i > k$. DGPs III, IV, V, and VI have $k = 1$, $k = 2$, $k = 3$, and $k = 6$, respectively. The true vectors of parameters in these DGPs are the estimates obtained by fitting VARMA(1,1) models to the first rolling window of Dataset 1 in their respective system dimensions. File "DGPsfile.csv" (available online) contains the true values. n accounts for the number of free parameters in A_0, A_1, and M_1 parameter matrices. MRRMSE is the mean of the RRMSE measures of all parameters. The lowest MRRMSE is highlighted in bold. RRMSE measures are computed as the ratio of the RMSE (root median squared error) measures obtained from a given estimator over the HR estimator. Share is the percentage over the total number of free parameters for which a given estimator delivers the lowest MRRMSE. The highest Share is highlighted in bold. Convergence is the percentage of replications in which the algorithms converged and yielded invertible and stable models. HR is the two-stage estimator of Hannan and Rissanen (1982); DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage estimator of Hannan and Kavalieris (1984); and KP is the multivariate version of the three-step estimator of Koreisha and Pukkila (1990) as formulated in Kascha (2012). The number of replications is set to 1,000.

Table 3: Forecast Summary: VARMA Out-of-Sample Performance Relative to Alternative Group of Models

$K = 3$

	VARMA					$\mathbf{p}_{k=1}$					$\mathbf{p}_{k=LP}$				
	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR
Hor: 1	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	**100%**	**100%**	**100%**	33%
Hor: 2	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	**100%**	**100%**	**100%**	**67%**	**67%**	**100%**	**100%**	33%
Hor: 3	**100%**	**67%**	**100%**	**100%**	**100%**	**67%**	33%	**67%**	**100%**	**67%**	**100%**	33%	**100%**	**100%**	**67%**
Hor: 6	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	**100%**	**100%**	**67%**	33%	33%	**67%**	**67%**
Hor: 9	**67%**	**100%**	**100%**	0%	**67%**	**67%**	**67%**	**100%**	0%	**67%**	0%	0%	0%	0%	0%
Hor: 12	**67%**	**67%**	**67%**	33%	**67%**	**67%**	33%	33%	33%	33%	33%	33%	0%	0%	33%

$K = 10$

	VARMA					$\mathbf{p}_{k=1}$					$\mathbf{p}_{LP}$				
	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR
Hor: 1	**83%**	**92%**	**100%**	**67%**	42%	**83%**	**83%**	**100%**	**58%**	42%	**50%**	**58%**	**92%**	25%	17%
Hor: 2	**100%**	**83%**	**100%**	**92%**	**100%**	**92%**	**75%**	**100%**	**83%**	**92%**	**67%**	**58%**	**92%**	**67%**	42%
Hor: 3	**100%**	**50%**	**100%**	**92%**	**100%**	**83%**	33%	**83%**	**75%**	**58%**	**50%**	25%	**75%**	42%	25%
Hor: 6	**92%**	**100%**	**92%**	**100%**	**100%**	**100%**	**83%**	**67%**	**100%**	**50%**	**58%**	**58%**	**50%**	**75%**	33%
Hor: 9	**100%**	**92%**	**58%**	33%	**83%**	**100%**	**92%**	**50%**	17%	**50%**	**50%**	**58%**	8%	25%	33%
Hor: 12	**100%**	**92%**	**67%**	**67%**	**92%**	**58%**	**58%**	**50%**	**58%**	**75%**	33%	33%	33%	17%	42%

$K = 20$

	VARMA					$\mathbf{p}_{k=1}$					$\mathbf{p}_{LP}$				
	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR
Hor: 1	**75%**	**67%**	**100%**	**75%**	**75%**	**67%**	**58%**	**92%**	**58%**	**67%**	**58%**	33%	**75%**	**50%**	33%
Hor: 2	**100%**	**75%**	**100%**	**75%**	**100%**	**75%**	**58%**	**92%**	**67%**	**83%**	**75%**	**50%**	**83%**	**50%**	**75%**
Hor: 3	**100%**	42%	**100%**	**67%**	**92%**	**83%**	33%	**58%**	17%	**50%**	**67%**	33%	**58%**	42%	**58%**
Hor: 6	**83%**	**100%**	**75%**	**100%**	**83%**	**100%**	**75%**	**58%**	**100%**	**50%**	**83%**	**67%**	**58%**	**67%**	**50%**
Hor: 9	**100%**	**83%**	**75%**	33%	**83%**	**100%**	**75%**	**67%**	33%	**58%**	**83%**	**75%**	**58%**	25%	**67%**
Hor: 12	**100%**	**75%**	**75%**	42%	**100%**	**58%**	**67%**	**58%**	42%	42%	42%	42%	33%	42%	25%

$K = 40$

	VARMA					$\mathbf{p}_{k=1}$					$\mathbf{p}_{LP}$				
	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR
Hor: 1	**83%**	**75%**	**100%**	**67%**	**92%**	**67%**	42%	**100%**	33%	**67%**	**67%**	**67%**	**100%**	**50%**	**75%**
Hor: 2	**83%**	**83%**	**100%**	**58%**	**92%**	**67%**	**50%**	**100%**	33%	**75%**	**75%**	**67%**	**100%**	**50%**	**75%**
Hor: 3	**100%**	**58%**	**100%**	**75%**	**92%**	42%	33%	**100%**	17%	42%	**67%**	25%	**100%**	**58%**	**58%**
Hor: 6	**92%**	**100%**	**75%**	**56%**	**67%**	**100%**	**67%**	**100%**	42%	42%	**67%**	**58%**	**100%**	42%	**58%**
Hor: 9	**100%**	**83%**	**100%**	33%	**75%**	**100%**	**67%**	**100%**	8%	33%	**58%**	42%	**100%**	25%	42%
Hor: 12	**92%**	**83%**	**100%**	**58%**	**75%**	33%	**67%**	**100%**	42%	42%	42%	**58%**	**100%**	**58%**	**58%**

$K = 52$

	VARMA					$\mathbf{p}_{k=1}$					$\mathbf{p}_{LP}$				
	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR	AR	ARMA	VAR	FM	BVAR
Hor: 1	**67%**	33%	**100%**	33%	**67%**	**67%**	33%	**100%**	33%	**67%**	0%	0%	**100%**	0%	**67%**
Hor: 2	**67%**	33%	**100%**	33%	**67%**	**67%**	33%	**100%**	33%	**67%**	33%	33%	**100%**	33%	33%
Hor: 3	**100%**	**67%**	**100%**	33%	**100%**	**67%**	33%	**100%**	0%	**67%**	0%	0%	**100%**	0%	0%
Hor: 6	**100%**	**67%**	**100%**	**67%**	**67%**	**100%**	**67%**	**100%**	**67%**	33%	0%	0%	**100%**	0%	0%
Hor: 9	**100%**	**67%**	**100%**	33%	**67%**	**100%**	**67%**	**100%**	0%	33%	0%	0%	**100%**	0%	0%
Hor: 12	**67%**	**100%**	**100%**	**67%**	**67%**	33%	**67%**	**100%**	33%	33%	33%	33%	**100%**	**67%**	33%

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9, and Hor: 12 account for one-, two- three- six-, nine-, and twelve-month-ahead forecast, respectively. For each system size, the first panel reports the frequency (in percentage points) for which at least one of the VARMA specifications ($\mathbf{p}_{k=1}$, $\mathbf{p}_{k=2}$, $\mathbf{p}_{k=3}$, $\mathbf{p}_{k=4}$, $\mathbf{p}_{k=5}$, $\mathbf{p}_{k=6}$, $\mathbf{p}_{LP}$, and $\mathbf{p}_{HK}$) outperforms (delivers the lowest RelMSFE measures) the assigned group of competitors in a given forecast horizon. The second and third panel report the frequencies for which the $\mathbf{p}_{k=1}$ and $\mathbf{p}_{LP}$ specifications, respectively, outperform the assigned group of competitors. We consider five groups of competitors. AR collects the AR(1) model; ARMA has the ARMA(1,1) specifications estimated with the IOLS and MLE estimators, VAR contains the VAR(p^*) model, where p^* is obtained by minimizing the AIC criterion; FM gathers the factor model specifications, namely FM_{IC_3} and FM_{SC}; and BVAR aggregates the three Bayesian VAR models: BVAR_{SC}, $\text{BVAR}_{0.2}$, and BVAR_{opt}. Percentages are also computed across the four datasets discussed in Section 5.1. Values greater or equal than 50% are highlighted in bold.

Table 4: Forecast: IOLS Out-of-Sample Performance Relative to Alternative VARMA Estimators

$K = 3$

	VARMA			$\mathbf{p}_{k=1}$			$\mathbf{p}_{k=2}$			$\mathbf{p}_{k=3}$			$\mathbf{p}_{k=4}$			$\mathbf{p}_{k=5}$			$\mathbf{p}_{k=6}$			$\mathbf{p}_{k=LP}$			$\mathbf{p}_{k=HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP										DJ2	HK	KP	DJ2	HK	KP
Hor: 1	**73%**	**80%**	**67%**	**100%**	33%	**100%**	**100%**	**67%**	**67%**	33%	**100%**	33%	-	-	-	-	-	-	-	-	-	33%	**100%**	33%	**100%**	**100%**	**100%**
Hor: 2	**87%**	**87%**	**60%**	**100%**	**67%**	**100%**	**100%**	**67%**	**67%**	**67%**	**100%**	33%	-	-	-	-	-	-	-	-	-	**67%**	**100%**	33%	**100%**	**100%**	**67%**
Hor: 3	**73%**	**67%**	**53%**	**100%**	**67%**	**67%**	**100%**	**67%**	**67%**	33%	**67%**	33%	-	-	-	-	-	-	-	-	-	33%	**67%**	33%	**100%**	**67%**	**67%**
Hor: 6	**73%**	**80%**	**67%**	**100%**	**100%**	**100%**	**100%**	0%	**100%**	33%	**100%**	33%	-	-	-	-	-	-	-	-	-	33%	**100%**	33%	**100%**	**100%**	**67%**
Hor: 9	**60%**	**87%**	40%	**100%**	**100%**	**67%**	**100%**	33%	**67%**	0%	**100%**	0%	-	-	-	-	-	-	-	-	-	0%	**100%**	0%	**100%**	**100%**	**67%**
Hor: 12	**73%**	**80%**	47%	**100%**	**100%**	33%	**100%**	**100%**	**67%**	33%	**67%**	33%	-	-	-	-	-	-	-	-	-	33%	**67%**	33%	**100%**	**67%**	**67%**

$K = 10$

	VARMA			$\mathbf{p}_{k=1}$			$\mathbf{p}_{k=2}$			$\mathbf{p}_{k=3}$			$\mathbf{p}_{k=4}$			$\mathbf{p}_{k=5}$			$\mathbf{p}_{k=6}$			$\mathbf{p}_{k=LP}$			$\mathbf{p}_{k=HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP
Hor: 1	45%	**50%**	40%	**75%**	**58%**	42%	42%	**58%**	**50%**	**58%**	42%	**50%**	**50%**	**50%**	**58%**	**50%**	**50%**	33%	25%	33%	**50%**	33%	42%	17%	25%	**67%**	17%
Hor: 2	42%	41%	47%	**100%**	**75%**	**92%**	**92%**	**83%**	**92%**	42%	33%	**50%**	25%	42%	**50%**	17%	25%	25%	8%	17%	25%	25%	25%	8%	25%	25%	33%
Hor: 3	**60%**	46%	**56%**	**100%**	**83%**	**83%**	**100%**	**67%**	**83%**	**92%**	**58%**	**67%**	42%	42%	**83%**	33%	42%	**58%**	33%	8%	33%	25%	**50%**	25%	**58%**	17%	17%
Hor: 6	**64%**	49%	**64%**	**100%**	**75%**	**100%**	**100%**	**58%**	**75%**	**67%**	33%	**67%**	42%	33%	**67%**	33%	**50%**	**58%**	**75%**	42%	**83%**	25%	**58%**	25%	**67%**	42%	33%
Hor: 9	**60%**	**56%**	**60%**	**100%**	**83%**	**100%**	**100%**	**75%**	**100%**	42%	**58%**	42%	25%	**50%**	**50%**	42%	33%	**50%**	**58%**	42%	**58%**	**50%**	**58%**	42%	**67%**	**50%**	42%
Hor: 12	**77%**	**61%**	**73%**	**75%**	**92%**	**67%**	**92%**	**92%**	**75%**	**75%**	**75%**	**75%**	**67%**	**67%**	**75%**	**67%**	33%	**83%**	**75%**	33%	**75%**	**75%**	**50%**	**75%**	**92%**	**50%**	**58%**

$K = 20$

	VARMA			$\mathbf{p}_{k=1}$			$\mathbf{p}_{k=2}$			$\mathbf{p}_{k=3}$			$\mathbf{p}_{k=4}$			$\mathbf{p}_{k=5}$			$\mathbf{p}_{k=6}$			$\mathbf{p}_{k=LP}$			$\mathbf{p}_{k=HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP
Hor: 1	45%	40%	**66%**	**50%**	0%	42%	33%	**50%**	**58%**	**67%**	**58%**	**75%**	**58%**	33%	**75%**	**58%**	**58%**	**92%**	**50%**	**83%**	**92%**	8%	8%	**75%**	33%	25%	17%
Hor: 2	43%	34%	**57%**	**83%**	17%	42%	**75%**	**50%**	**75%**	42%	25%	**75%**	33%	33%	**50%**	33%	33%	**92%**	17%	33%	**67%**	17%	42%	**58%**	42%	42%	0%
Hor: 3	**64%**	48%	**61%**	**100%**	42%	33%	**100%**	**58%**	42%	**58%**	**50%**	**67%**	**75%**	**58%**	**75%**	**58%**	42%	**100%**	42%	42%	**50%**	25%	42%	**67%**	**50%**	**50%**	**58%**
Hor: 6	**70%**	31%	**61%**	**100%**	42%	**75%**	**100%**	33%	**75%**	**100%**	25%	**58%**	**75%**	25%	**67%**	42%	17%	**92%**	**50%**	33%	**50%**	42%	**50%**	**58%**	**50%**	25%	17%
Hor: 9	**65%**	**65%**	**72%**	**100%**	**92%**	**92%**	**100%**	**92%**	**92%**	**75%**	**58%**	**58%**	**50%**	**58%**	**67%**	33%	**50%**	**75%**	33%	**75%**	**67%**	**58%**	**50%**	**83%**	**67%**	42%	42%
Hor: 12	**74%**	**64%**	**76%**	**83%**	**100%**	**75%**	**83%**	**92%**	**83%**	**92%**	**50%**	**75%**	**67%**	42%	**83%**	**50%**	33%	**92%**	**83%**	**50%**	**83%**	**67%**	**58%**	**92%**	**67%**	**83%**	25%

$K = 40$

	VARMA			$\mathbf{p}_{k=1}$			$\mathbf{p}_{k=2}$			$\mathbf{p}_{k=3}$			$\mathbf{p}_{k=4}$			$\mathbf{p}_{k=5}$			$\mathbf{p}_{k=6}$			$\mathbf{p}_{k=LP}$			$\mathbf{p}_{k=HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP
Hor: 1	**85%**	**81%**	**98%**	**58%**	**75%**	**92%**	**100%**	**92%**	**100%**	**100%**	**100%**	**100%**	**100%**	**92%**	**100%**	**100%**	**67%**	**100%**	**100%**	**92%**	**100%**	**75%**	**67%**	**92%**	**50%**	**67%**	**100%**
Hor: 2	**76%**	**85%**	**97%**	**67%**	**100%**	**100%**	**92%**	**100%**	**100%**	**92%**	**92%**	**100%**	**100%**	**92%**	**100%**	**75%**	**75%**	**100%**	**83%**	**75%**	**100%**	**58%**	**58%**	**83%**	42%	**92%**	**92%**
Hor: 3	**72%**	**91%**	**93%**	**92%**	**100%**	**92%**	**100%**	**100%**	**100%**	**92%**	**100%**	**100%**	**100%**	**100%**	**100%**	42%	**92%**	**75%**	**75%**	**92%**	**92%**	**50%**	**58%**	**100%**	25%	**83%**	**83%**
Hor: 6	**68%**	**82%**	**81%**	**100%**	**75%**	**92%**	**100%**	**92%**	**75%**	**83%**	**92%**	**75%**	**100%**	**92%**	**92%**	8%	**75%**	**58%**	33%	**83%**	**67%**	**50%**	**58%**	**100%**	**67%**	**92%**	**92%**
Hor: 9	**73%**	**79%**	**88%**	**100%**	**75%**	**92%**	**100%**	**75%**	**67%**	**83%**	**83%**	**100%**	**83%**	**83%**	**92%**	25%	**67%**	**75%**	**50%**	**83%**	**75%**	**75%**	**92%**	**100%**	**67%**	**75%**	**100%**
Hor: 12	**73%**	**86%**	**68%**	**100%**	**100%**	**58%**	**100%**	**83%**	42%	**100%**	**100%**	**58%**	**50%**	**92%**	**75%**	33%	**67%**	**67%**	42%	**75%**	**75%**	**100%**	**92%**	**92%**	**58%**	**83%**	**75%**

$K = 52$

	VARMA			$\mathbf{p}_{k=1}$			$\mathbf{p}_{k=2}$			$\mathbf{p}_{k=3}$			$\mathbf{p}_{k=4}$			$\mathbf{p}_{k=5}$			$\mathbf{p}_{k=6}$			$\mathbf{p}_{k=LP}$			$\mathbf{p}_{k=HK}$		
	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP	DJ2	HK	KP
Hor: 1	**83%**	**83%**	**96%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	0%	**100%**	0%	**67%**	**67%**
Hor: 2	**58%**	**83%**	**88%**	**67%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	**100%**	**100%**	**67%**	**100%**	**100%**	**67%**	**100%**	**100%**	**100%**	**100%**	**100%**	0%	0%	33%	0%	**67%**	**67%**
Hor: 3	**75%**	**83%**	**88%**	**67%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	0%	0%	33%	33%	**67%**	**67%**
Hor: 6	**83%**	**88%**	**96%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	0%	0%	**100%**	**67%**	**100%**	**100%**
Hor: 9	**67%**	**92%**	**88%**	**100%**	**100%**	**100%**	**67%**	**100%**	**67%**	**67%**	**100%**	**100%**	**67%**	**100%**	**100%**	**67%**	**100%**	**100%**	**67%**	**100%**	33%	0%	33%	**100%**	**100%**	**100%**	**100%**
Hor: 12	**50%**	**63%**	**67%**	33%	**67%**	**100%**	**67%**	**67%**	**100%**	33%	**100%**	**100%**	**67%**	**100%**	**100%**	**67%**	**100%**	**100%**	**67%**	**100%**	33%	**67%**	**67%**	**100%**	**67%**	33%	33%

Hor:1, Hor: 2, Hor: 3, Hor: 6, Hor: 9, and Hor: 12 account for one-, two- three- six-, nine-, and twelve-month-ahead forecast, respectively. The first set of results, denoted as VARMA, summarizes the relative forecast performance of all VARMA specifications. The remaining panels summarize the results for the eight VARMA specifications, i.e. $\mathbf{p}_{k=1}$, $\mathbf{p}_{k=2}$, $\mathbf{p}_{k=3}$, $\mathbf{p}_{k=4}$, $\mathbf{p}_{k=5}$, $\mathbf{p}_{k=6}$, $\mathbf{p}_{LP}$, and $\mathbf{p}_{HK}$. We report the frequency (in percentage points) which the IOLS estimator delivers lower RelMSFE measures than the assigned competitor. Percentages are also computed across the four datasets. DJ2 is the two-step estimator of Dufour and Jouini (2014); HK is the three-stage estimator of Hannan and Kavalieris (1984) and KP is the three-step multivariate version of Koreisha and Pukkila (1990). Values greater or equal than 50% are highlighted in bold.